



US009047351B2

(12) **United States Patent**
Riddle et al.

(10) **Patent No.:** **US 9,047,351 B2**
(45) **Date of Patent:** **Jun. 2, 2015**

(54) **CLUSTER OF PROCESSING NODES WITH DISTRIBUTED GLOBAL FLASH MEMORY USING COMMODITY SERVER TECHNOLOGY**

(75) Inventors: **Thomas A. Riddle**, Belmont, CA (US); **Darpan Dinker**, Union City, CA (US); **Andrew D. Eckhardt**, Menlo Park, CA (US); **Michael J. Koster**, Bridgeville, CA (US)

(73) Assignee: **SANDISK ENTERPRISE IP LLC**, Milpitas, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 909 days.

(21) Appl. No.: **13/084,502**

(22) Filed: **Apr. 11, 2011**

(65) **Prior Publication Data**

US 2012/0017037 A1 Jan. 19, 2012

Related U.S. Application Data

(60) Provisional application No. 61/323,302, filed on Apr. 12, 2010, provisional application No. 61/323,351, filed on Apr. 12, 2010.

(51) **Int. Cl.**
G06F 17/30 (2006.01)

(52) **U.S. Cl.**
CPC **G06F 17/30557** (2013.01); **G06F 17/30519** (2013.01)

(58) **Field of Classification Search**
CPC G06F 3/067; G06F 12/0246; G06F 12/00; G06F 13/00; G06F 12/0831; G06F 12/0817; G06F 12/0813; G06F 12/0815; G06F 12/0811
USPC 711/103, E12.008, 115, 130, 141–146, 711/147–148, 153

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,916,605 A 4/1990 Beardsley et al.
5,046,002 A 9/1991 Takashi et al.

(Continued)

FOREIGN PATENT DOCUMENTS

EP 1548600 B1 1/2007
EP 1746510 A1 1/2007

OTHER PUBLICATIONS

Mukherjee, R., Nakayama, Y., Mima, T. "Verification of an industrial CC-NUMA server." Jan. 7, 2002-Jan. 11, 2002. Design Automation Conference, 2002. Proceedings of ASP-DAC 2002. 7th Asia and South Pacific and the 15th International Conference on VLSI Design. pp. 747-752.*

(Continued)

Primary Examiner — Mohammad Islam

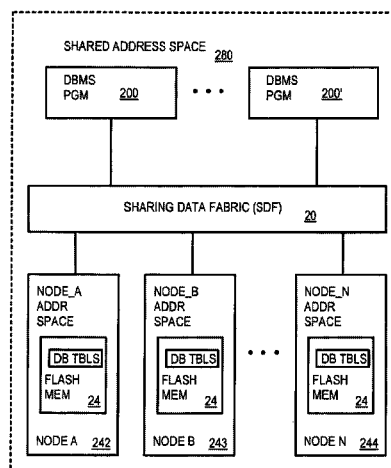
Assistant Examiner — Mikka Liu

(74) *Attorney, Agent, or Firm* — Morgan, Lewis & Bockius LLP

(57) **ABSTRACT**

Approaches for a distributed storage system that comprises a plurality of nodes. Each node, of the plurality of nodes, executes one or more application processes which are capable of accessing persistent shared memory. The persistent shared memory is implemented by solid state devices physically maintained on each of the plurality of nodes. Each the one or more application processes, maintained on a particular node, of the plurality of nodes, communicates with a shared data fabric (SDF) to access the persistent shared memory. The persistent shared memory comprises a scoreboard implemented in shared DRAM memory that is mapped to a persistent storage. The scoreboard provides a crash tolerant mechanism for enabling application processes to communicate with the shared data fabric (SDF).

11 Claims, 38 Drawing Sheets



(56)

References Cited

U.S. PATENT DOCUMENTS

5,057,996 A	10/1991	Cutler et al.	7,885,923 B1 *	2/2011	Tawri et al. 707/610
5,117,350 A	5/1992	Parrish et al.	7,917,472 B2	3/2011	Persson
5,212,789 A	5/1993	Rago	8,015,352 B2	9/2011	Zhang et al.
5,287,496 A	2/1994	Chen et al.	8,018,729 B2	9/2011	Skinner
5,297,258 A	3/1994	Hale et al.	8,024,515 B2	9/2011	Auerbach et al.
5,394,555 A	2/1995	Hunter et al.	8,037,349 B2	10/2011	Mandagere et al.
5,403,639 A	4/1995	Belsan et al.	8,069,328 B2	11/2011	Pyeon
5,423,037 A	6/1995	Hvasshovd	8,099,391 B1	1/2012	Monckton
5,509,134 A	4/1996	Fandrich et al.	8,103,643 B2	1/2012	Danilov et al.
5,537,534 A	7/1996	Voigt et al.	8,205,206 B2	6/2012	Ozer et al.
5,603,001 A	2/1997	Sukegawa et al.	8,225,053 B1 *	7/2012	McCorkendale et al. 711/159
5,611,057 A	3/1997	Pecone et al.	8,239,617 B1	8/2012	Linnell
5,613,071 A	3/1997	Rankin et al.	8,261,266 B2	9/2012	Pike et al.
5,680,579 A	10/1997	Young et al.	8,261,289 B2	9/2012	Kasravi et al.
5,692,149 A	11/1997	Lee	8,321,450 B2	11/2012	Thatte et al.
5,701,480 A	12/1997	Raz	8,335,776 B2	12/2012	Gokhale
5,742,787 A	4/1998	Talreja	8,370,853 B2	2/2013	Giampaolo et al.
5,887,138 A	3/1999	Hagersten et al.	8,401,994 B2	3/2013	Hoang et al.
5,897,661 A	4/1999	Baranovsky et al.	8,504,526 B2	8/2013	Gokhale et al.
5,897,664 A	4/1999	Nesheim et al.	8,666,939 B2	3/2014	O'Kafka et al.
5,960,436 A	9/1999	Chang et al.	8,671,074 B2	3/2014	Wang et al.
5,963,983 A	10/1999	Sakakura et al.	2002/0089933 A1	7/2002	Giroux et al.
6,000,006 A	12/1999	Bruce et al.	2002/0129192 A1	9/2002	Spiegel et al.
6,023,745 A *	2/2000	Lu 711/5	2002/0166031 A1	11/2002	Chen et al.
6,052,815 A	4/2000	Zook	2002/0184239 A1	12/2002	Mosher, Jr. et al.
6,130,759 A	10/2000	Blair	2003/0016596 A1	1/2003	Chiquoine et al.
6,141,692 A	10/2000	Loewenstein et al.	2003/0097610 A1	5/2003	Hofner
6,151,688 A	11/2000	Wipfel et al.	2003/0177408 A1	9/2003	Fields et al.
6,216,126 B1	4/2001	Ronstrom	2003/0220985 A1	11/2003	Kawamoto et al.
6,298,390 B1	10/2001	Matena et al.	2004/0010502 A1	1/2004	Bomfim et al.
6,308,169 B1	10/2001	Ronstrom et al.	2004/0078379 A1	4/2004	Hinshaw et al.
6,434,144 B1	8/2002	Romanov	2004/0143562 A1	7/2004	Chen et al.
6,467,060 B1	10/2002	Malakapalli et al.	2004/0148283 A1	7/2004	Harris et al.
6,615,313 B2	9/2003	Kato et al.	2004/0172494 A1	9/2004	Petty et al.
6,658,526 B2	12/2003	Nguyen et al.	2004/0205151 A1	10/2004	Sprigg et al.
6,728,826 B2	4/2004	Kaki et al.	2004/0230862 A1	11/2004	Merchant et al.
6,745,209 B2	6/2004	Holenstein et al.	2004/0267835 A1	12/2004	Zwilling et al.
6,804,766 B1 *	10/2004	Noel et al. 711/203	2005/0005074 A1	1/2005	Landin et al.
6,874,044 B1	3/2005	Chou et al.	2005/0021565 A1	1/2005	Kapoor et al.
6,938,084 B2	8/2005	Gamache et al.	2005/0027701 A1	2/2005	Zane et al.
6,944,699 B1	9/2005	Bugnion et al.	2005/0028134 A1	2/2005	Zane et al.
6,981,070 B1	12/2005	Luk et al.	2005/0034048 A1	2/2005	Nemawarkar et al.
7,003,586 B1	2/2006	Bailey et al.	2005/0081091 A1	4/2005	Bartfai et al.
7,010,521 B2	3/2006	Hinshaw et al.	2005/0086413 A1	4/2005	Lee et al.
7,043,621 B2	5/2006	Merchant et al.	2005/0120133 A1	6/2005	Slack-Smith
7,082,481 B2	7/2006	Lambrache et al.	2005/0131964 A1	6/2005	Saxena
7,162,467 B2	1/2007	Eshleman et al.	2005/0240635 A1	10/2005	Kapoor et al.
7,200,718 B2	4/2007	Duzett	2005/0246487 A1	11/2005	Ergan et al.
7,203,890 B1	4/2007	Normoyle	2006/0059428 A1	3/2006	Humphries et al.
7,249,280 B2	7/2007	Lamport et al.	2006/0064549 A1	3/2006	Wintergerst
7,251,749 B1	7/2007	Fong et al.	2006/0085594 A1	4/2006	Roberson et al.
7,269,708 B2	9/2007	Ware	2006/0130063 A1	6/2006	Kilian et al.
7,269,755 B2	9/2007	Moshayedi et al.	2006/0161530 A1	7/2006	Biswal et al.
7,272,605 B1	9/2007	Hinshaw et al.	2006/0174063 A1	8/2006	Soules et al.
7,272,654 B1	9/2007	Brendel	2006/0174069 A1	8/2006	Shaw et al.
7,281,160 B2	10/2007	Stewart	2006/0179083 A1	8/2006	Kulkarni et al.
7,305,386 B2	12/2007	Hinshaw et al.	2006/0195648 A1	8/2006	Chandrasekaran et al.
7,334,154 B2	2/2008	Lorch et al.	2006/0212795 A1	9/2006	Cottrille et al.
7,359,927 B1	4/2008	Cardente	2006/0218210 A1	9/2006	Sarma et al.
7,383,290 B2	6/2008	Mehra et al.	2006/0242163 A1	10/2006	Miller et al.
7,406,487 B1	7/2008	Gupta et al.	2006/0253724 A1	11/2006	Zhang
7,415,488 B1 *	8/2008	Muth et al. 707/648	2007/0038794 A1	2/2007	Purcell et al.
7,417,992 B2	8/2008	Krishnan	2007/0043790 A1	2/2007	Kryger
7,467,265 B1	12/2008	Tawri et al.	2007/0043860 A1	2/2007	Pabari
7,529,882 B2	5/2009	Wong	2007/0073896 A1 *	3/2007	Rothman et al. 709/232
7,542,968 B2	6/2009	Yokomizo et al.	2007/0143368 A1	6/2007	Lundsgaard et al.
7,562,162 B2	7/2009	Kreiner et al.	2007/0156842 A1	7/2007	Vermuelen et al.
7,584,222 B1	9/2009	Georgiev	2007/0174541 A1	7/2007	Chandrasekaran et al.
7,610,445 B1	10/2009	Manus et al.	2007/0234182 A1	10/2007	Wickeraad et al.
7,623,494 B2	11/2009	Zhu et al.	2007/0276784 A1	11/2007	Piedmonte
7,627,618 B2	12/2009	Honigfort	2007/0283079 A1	12/2007	Iwamura et al.
7,647,449 B1	1/2010	Roy et al.	2007/0288692 A1	12/2007	Bruce et al.
7,657,710 B2 *	2/2010	Loewenstein 711/141	2007/0288792 A1	12/2007	Thorpe et al.
7,809,691 B1	10/2010	Karmarkar et al.	2007/0294564 A1	12/2007	Reddin et al.
7,822,711 B1	10/2010	Ranade	2007/0299816 A1	12/2007	Arora et al.
			2008/0016300 A1	1/2008	Yim et al.
			2008/0034076 A1	2/2008	Ishikawa et al.
			2008/0034174 A1	2/2008	Traister et al.
			2008/0034249 A1	2/2008	Husain et al.

(56)

References Cited**U.S. PATENT DOCUMENTS**

2008/0046538	A1	2/2008	Susarla et al.	
2008/0046638	A1	2/2008	Maheshwari et al.	
2008/0126706	A1	5/2008	Newport et al.	
2008/0172402	A1	7/2008	Birdwell et al.	
2008/0256103	A1	10/2008	Fachan et al.	
2008/0288713	A1	11/2008	Lee et al.	
2008/0288819	A1*	11/2008	Heller, Jr.	714/19
2008/0295105	A1	11/2008	Ozer et al.	
2008/0301256	A1*	12/2008	McWilliams et al.	709/214
2009/0006500	A1	1/2009	Shiozawa et al.	
2009/0006681	A1	1/2009	Hubert et al.	
2009/0006888	A1	1/2009	Bernhard et al.	
2009/0019456	A1	1/2009	Saxena et al.	
2009/0024871	A1	1/2009	Emaru et al.	
2009/0030943	A1	1/2009	Kall	
2009/0059539	A1	3/2009	Ryu et al.	
2009/0070530	A1	3/2009	Satoyama et al.	
2009/0150599	A1	6/2009	Bennett	
2009/0177666	A1	7/2009	Kaneda	
2009/0198791	A1	8/2009	Menghnani	
2009/0240664	A1*	9/2009	Dinker et al.	707/3
2009/0240869	A1*	9/2009	O’Kafka et al.	711/103
2009/0327751	A1	12/2009	Koifman et al.	
2010/0058021	A1*	3/2010	Kawamura	711/171
2010/0107017	A1*	4/2010	Munjal et al.	714/49
2010/0125695	A1	5/2010	Wu et al.	
2010/0241895	A1	9/2010	Li et al.	
2010/0262762	A1	10/2010	Borchers et al.	
2010/0299490	A1*	11/2010	Attarde et al.	711/162
2010/0318821	A1	12/2010	Kwan et al.	
2010/0325498	A1	12/2010	Nagadomi	
2011/0022566	A1	1/2011	Beaverson et al.	
2011/0072206	A1	3/2011	Ross et al.	
2011/0082985	A1	4/2011	Haines et al.	
2011/0099420	A1	4/2011	MacDonald McAlister	
2011/0167038	A1	7/2011	Wang et al.	
2011/0179279	A1	7/2011	Greevenbosch et al.	
2011/0185147	A1	7/2011	Hatfield et al.	
2011/0191299	A1	8/2011	Huynh Huu et al.	
2011/0225214	A1*	9/2011	Guo	707/813
2012/0005154	A1	1/2012	George et al.	
2013/0066948	A1	3/2013	Colrain et al.	

OTHER PUBLICATIONS

Shacham, O., Wachs, M., Solomatnikov, A., Firoozshahian, A., Richardson, S., Horowitz, M. “Verification of chip multiprocessor memory systems using a relaxed scoreboard.” Nov. 8-12, 2008. Microarchitecture, 2008. MICRO-41. 2008 41st IEEE/ACM International Symposium. pp. 294-305.*

bsn-modulestore, Versioning Concept, Oct. 13, 2010, 2 pgs.

Btrfs, <http://en.wikipedia.org>, Oct. 3, 2011, 9 pgs.

Buchholz, The Structure of the Reiser File System, Jan. 26, 2006, 21 pgs.

Chacon, Git, The Fast Version Control System, Oct. 3, 2011, 3 pgs.

Email Communication from James Bodwin to Christopher Brokaw re prior art, Sep. 13, 2011, 4 pgs.

Git (Software), <http://en.wikipedia.org>, Oct. 3, 2011, 10 pgs.

Hitz, File System Design for an NFS File Server Appliance, Jan. 19, 1994, 23 pgs.

McDonald, Architectural Semantics for Practical Transactional Memory, Jun. 2006, 12 pgs.

McGonigle, A Short History of btrfs, Aug. 14, 2009, 11 pgs.

Mellor, ZFS—the future of file systems? Aug. 14, 2006, 5 pgs.

Mercurial, <http://en.wikipedia.org>, Oct. 2, 2011, 6 pages.

Module: Mongoid: Versioning, <http://rdoc.info>, Documentation by YARD 0.7.2, 6 pages Oct. 3, 2011.

Noach, Database Schema under Version Control, code.openarck.org, Apr. 22, 2010, 6 pages.

Reiser FS, , <http://en.wikipedia.org>, Sep. 17, 2011, 5 pgs.

Rice, Extension Versioning, Update and Compatibility, Aug. 9, 2011, 11 pgs.

Rice, Toolkit Version Format, Aug. 19, 2011, 4 pgs.

Russell, Track and Record Database Schema Versions, Jun. 28, 2005, 8 pgs.

Schooner Information Technology, IPAF, PCT/US2008/065167, Oct. 23, 2008, 7 pgs.

Schooner Information Technology, ISR/WO, PCT/US2008/065167, Jan. 28, 2009, 16 pgs.

SQL Server Database Schema Versioning and Update, Dec. 2, 2009, 2 pgs.

Sufficiently Advanced Bug, File Versioning, Caching and Hashing, Oct. 3, 2011, 3 pgs.

The Z File System (ZFS), FreeBSD Handbook, Oct. 3, 2011, 8 pgs (Author not provided).

Tux3 Linux Filesystem Project, 2008, 1 pg.

Tux3 Versioning Filesystem, Jul. 2008, 67 pgs.

Tux3, <http://en.wikipedia.org>, Jun. 2, 2010, 3 pgs.

Vijaykumar, Speculative Versioning Cache, Dec. 1, 2001, 13 pgs.

WAFL—Write Anywhere File Layout, 1999, 1 pg.

Write Anywhere File Layout, Sep. 9, 2011, 2 pgs.

ZFS, , <http://en.wikipedia.org> Sep. 30, 2011, 18 pgs.

Ajmani, Automatic Software Upgrades for Distributed Systems, MIT, Sep. 2004, 164 pgs.

Amza, Data Replication Strategies for Fault Tolerance and Availability on Commodity Clusters, 2000, 9 pgs.

Chockler, Active Disk Paxos with infinitely many processes, Springer-Verlag, Apr. 2005, 12 pgs.

Dwork, Consensus in the presence of partial synchrony, MIT, 1988, 6 pgs.

Guerraoui, A Leader Election Protocol for Eventually Synchronous Shared Memory Systems, IEEE, 2006, 6 pgs.

Lamport, Cheap Paxos, Microsoft, 2004, 9 pgs.

Lamport, Fast Paxos, Microsoft, Jul. 2005, 43 pgs.

Lamport, Generalized Consensus and Paxos, Microsoft, Mar. 2004, 25 pgs.

Lamport, Paxos Made Simple, Nov. 2001, 14 pgs.

Malkhi, Lecture notes in computer science [Section: Omega Meets Paxos, Leader election and stability without eventual timely links], 2005, pp. 199-213.

Pease, Reaching Agreement in the Presence of Faults, ACM, 1980, pp. 228-234.

Schneider, Implementing fault tolerant services using the state machine, Cornell Univ., 1990, 21 pgs.

Unknown Author, Supermicro, “Intel Itanium Processor 9300 Series Based Server Systems,” Jul. 8, 2010, <http://www.supermicro.com/products/nfo/itanium.cfm>, 3 pages.

Walker, Hash Table Tutorial, Oct. 13, 2007, 14 pgs.

* cited by examiner

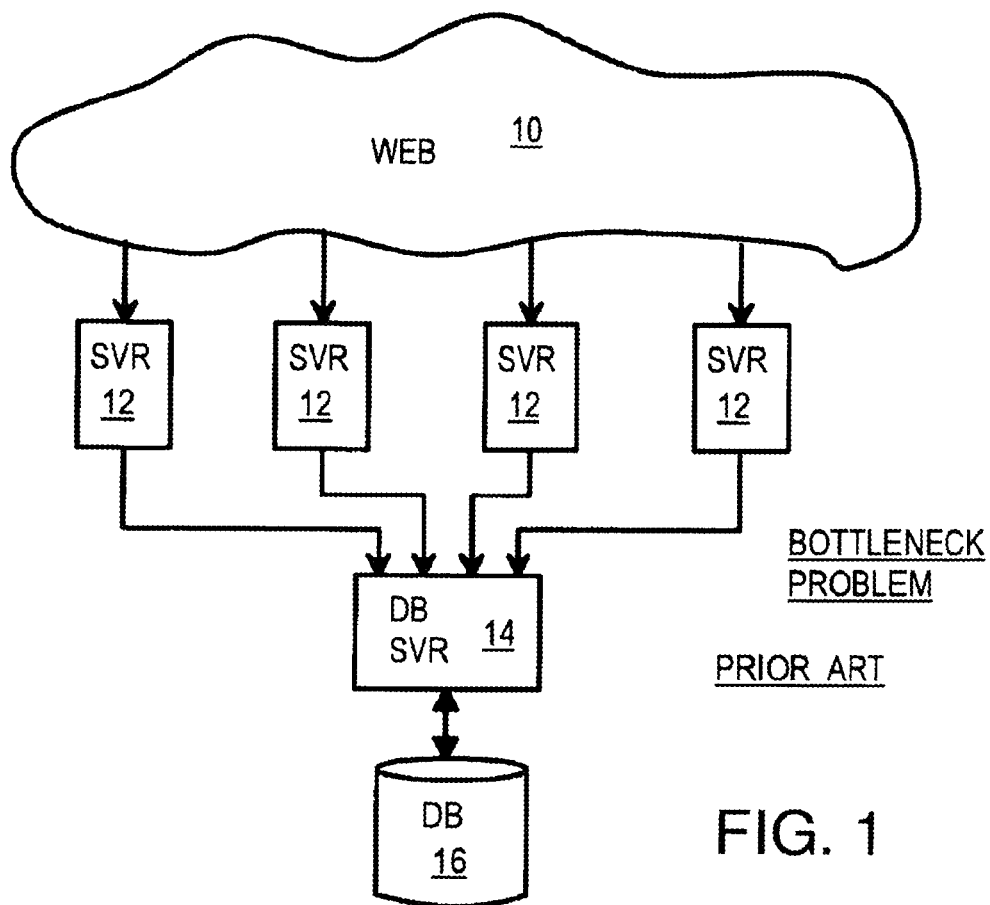


FIG. 1

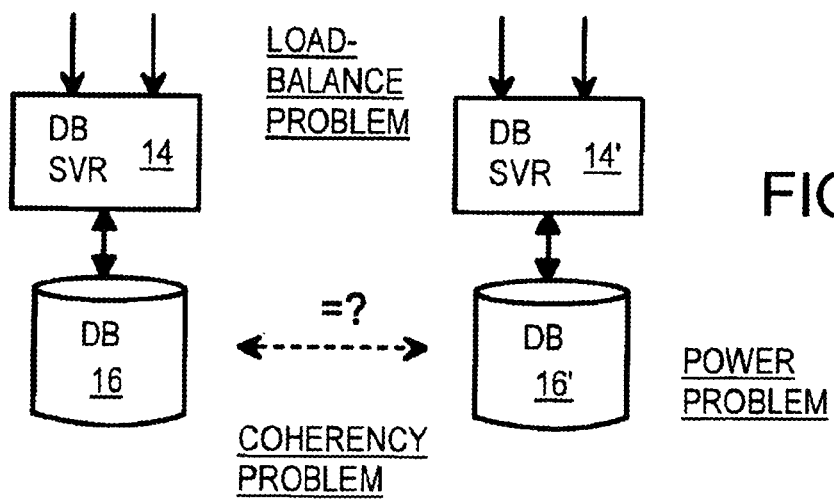


FIG. 2

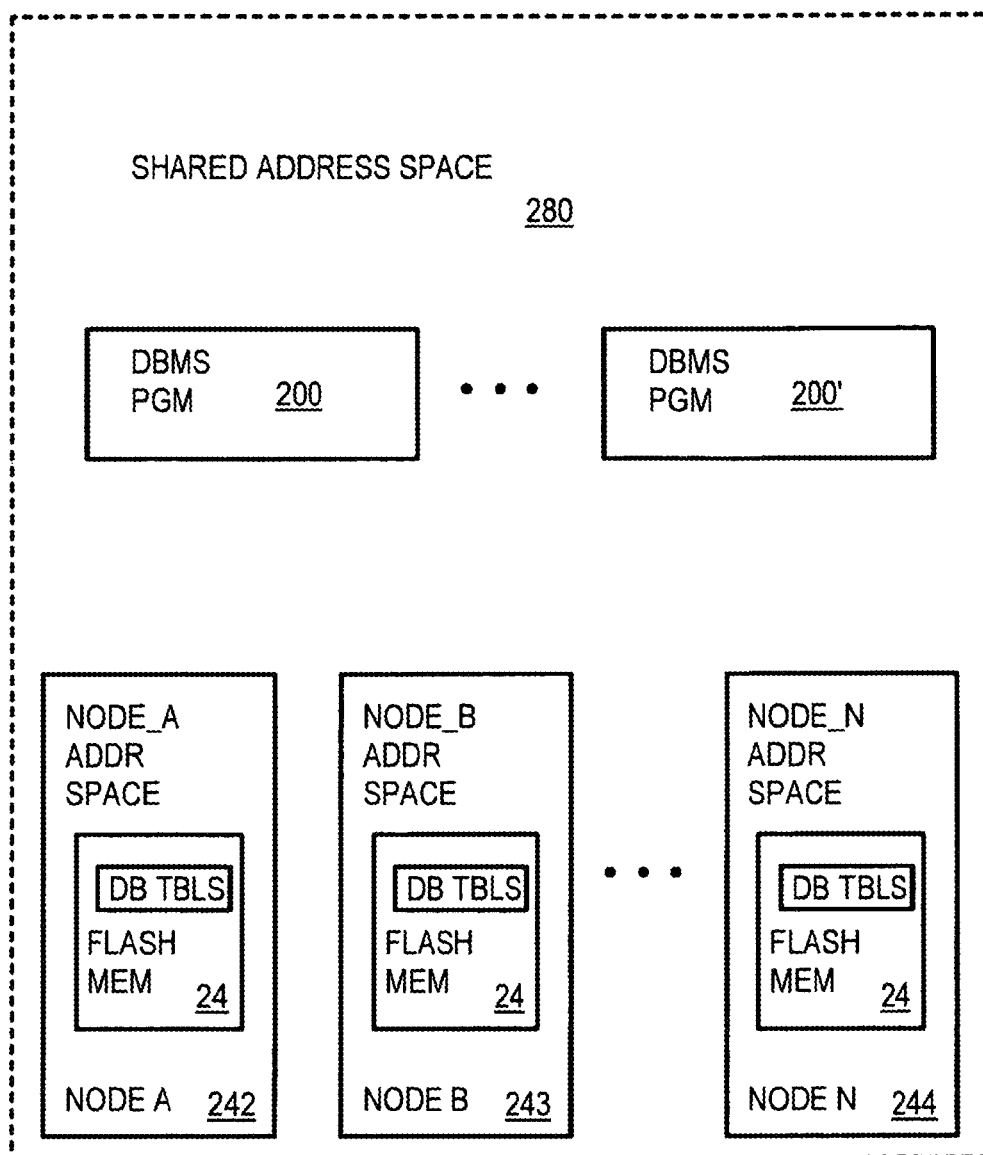


FIG. 3

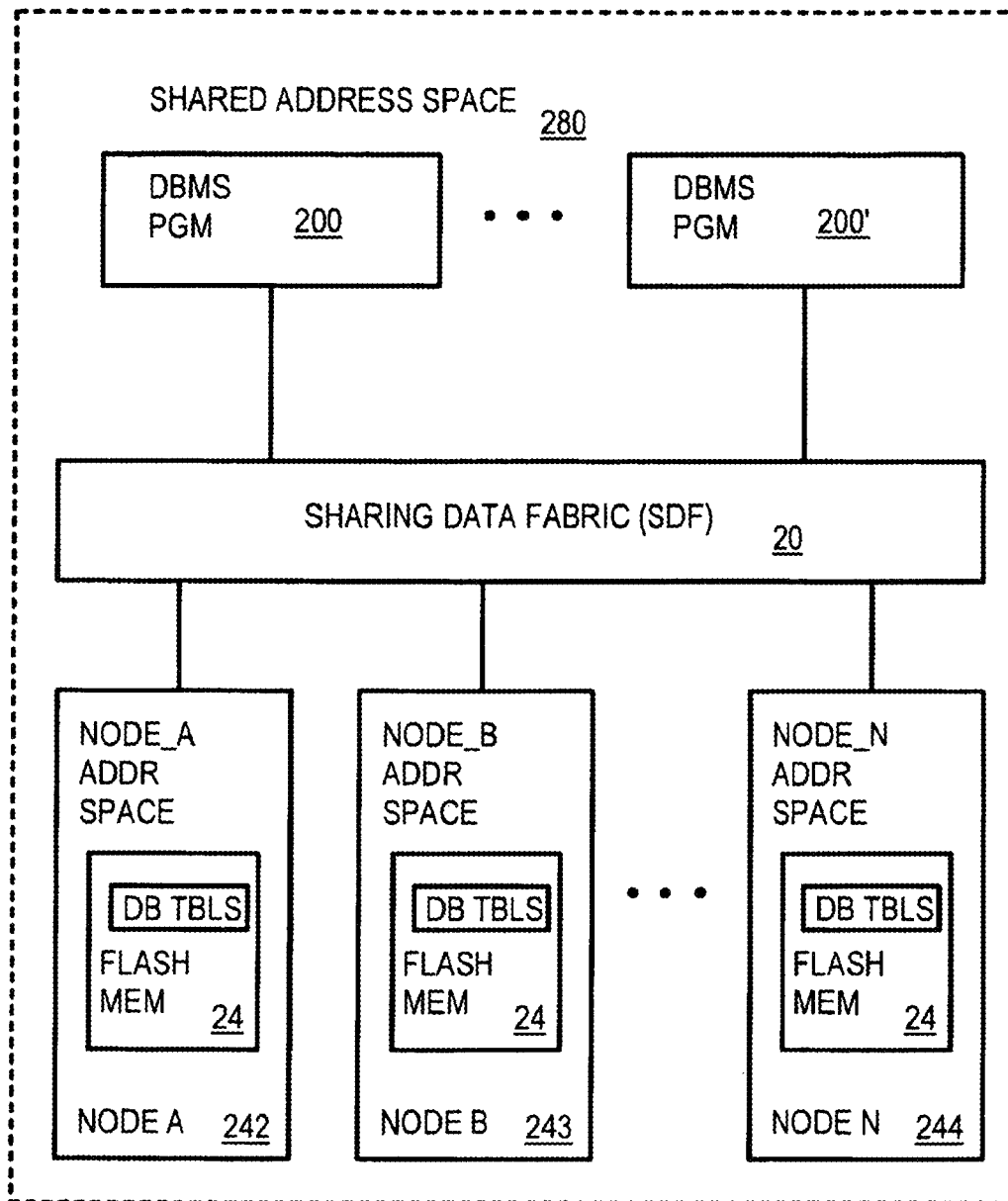


FIG. 4

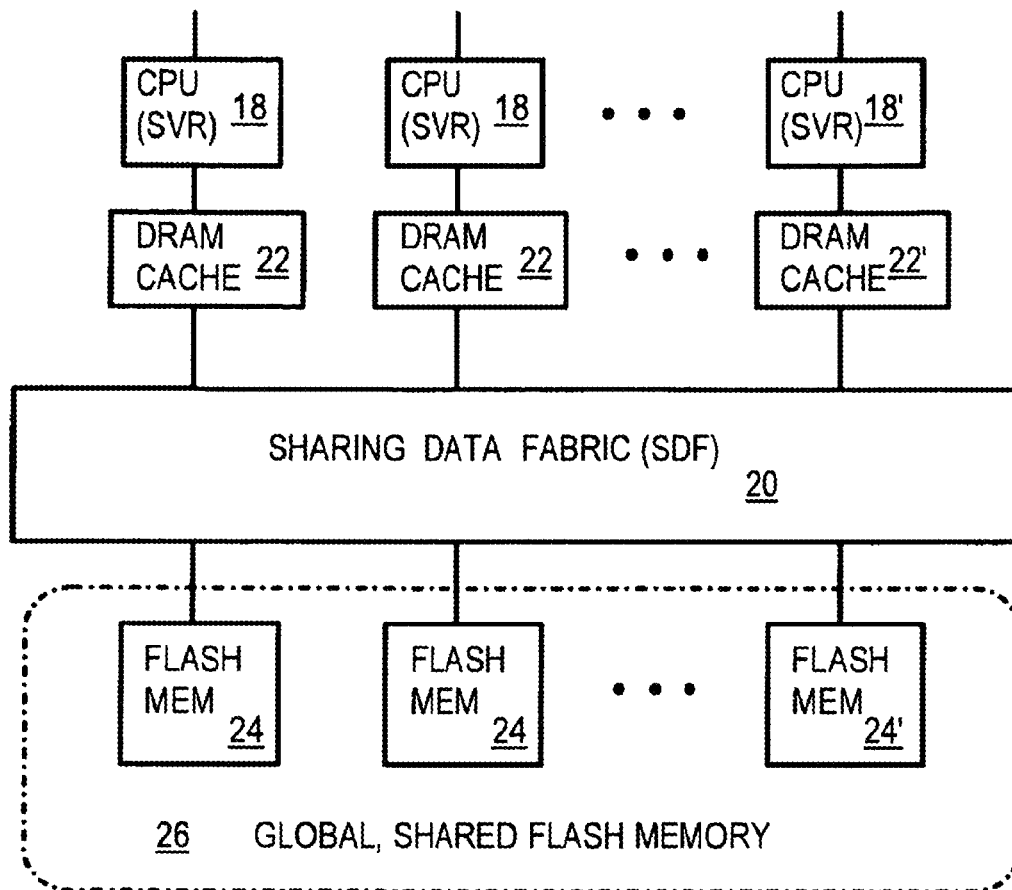
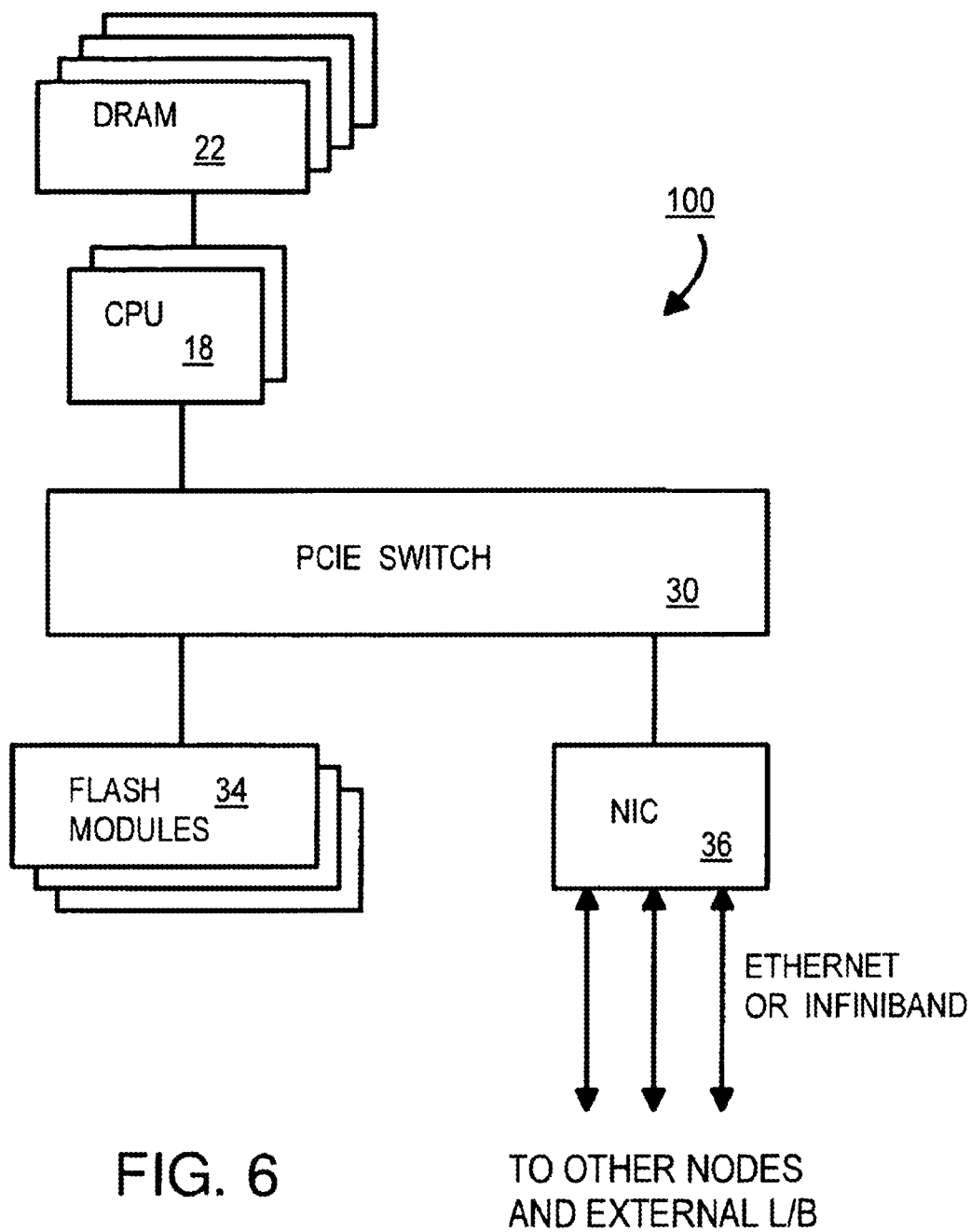


FIG. 5



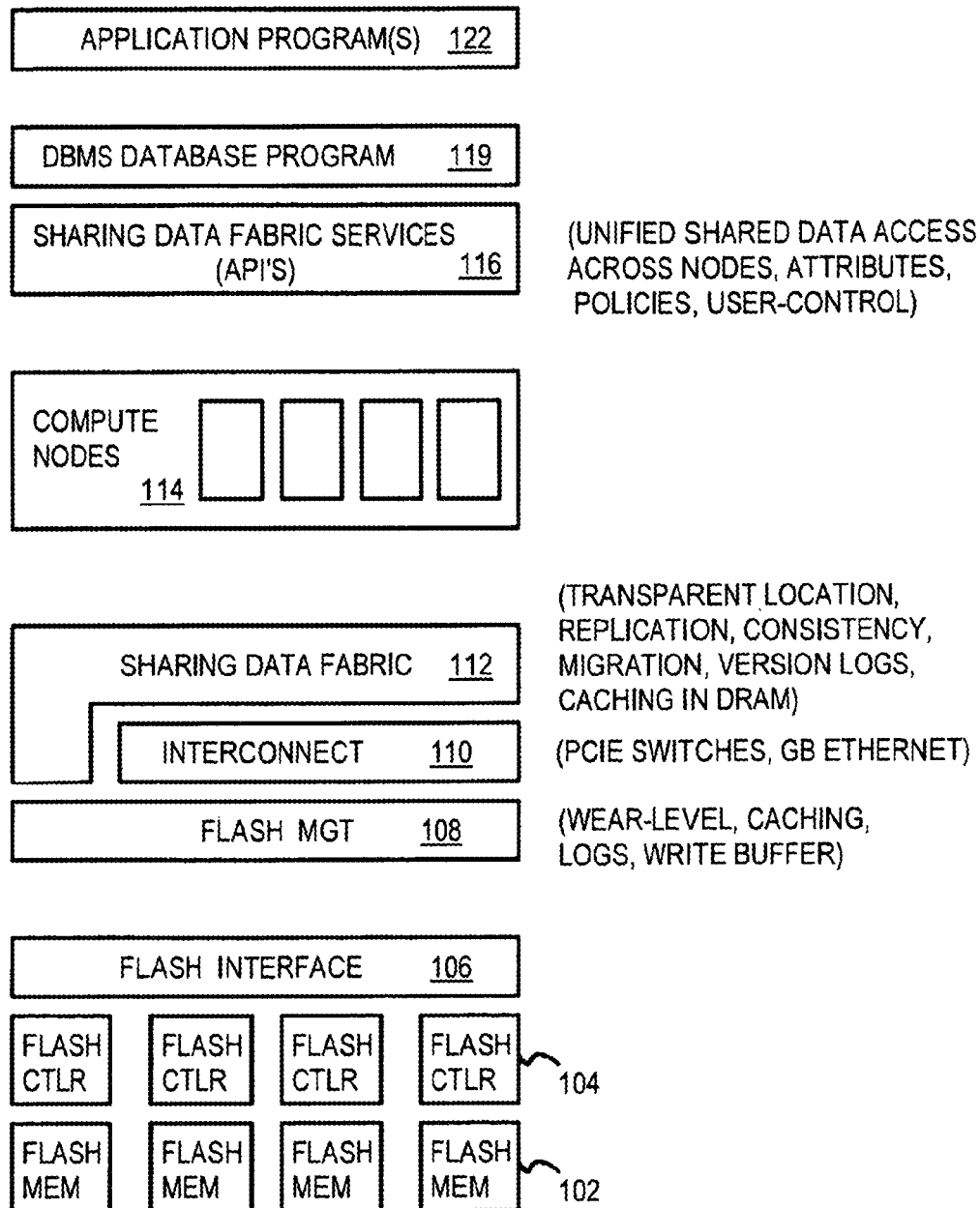


FIG. 7

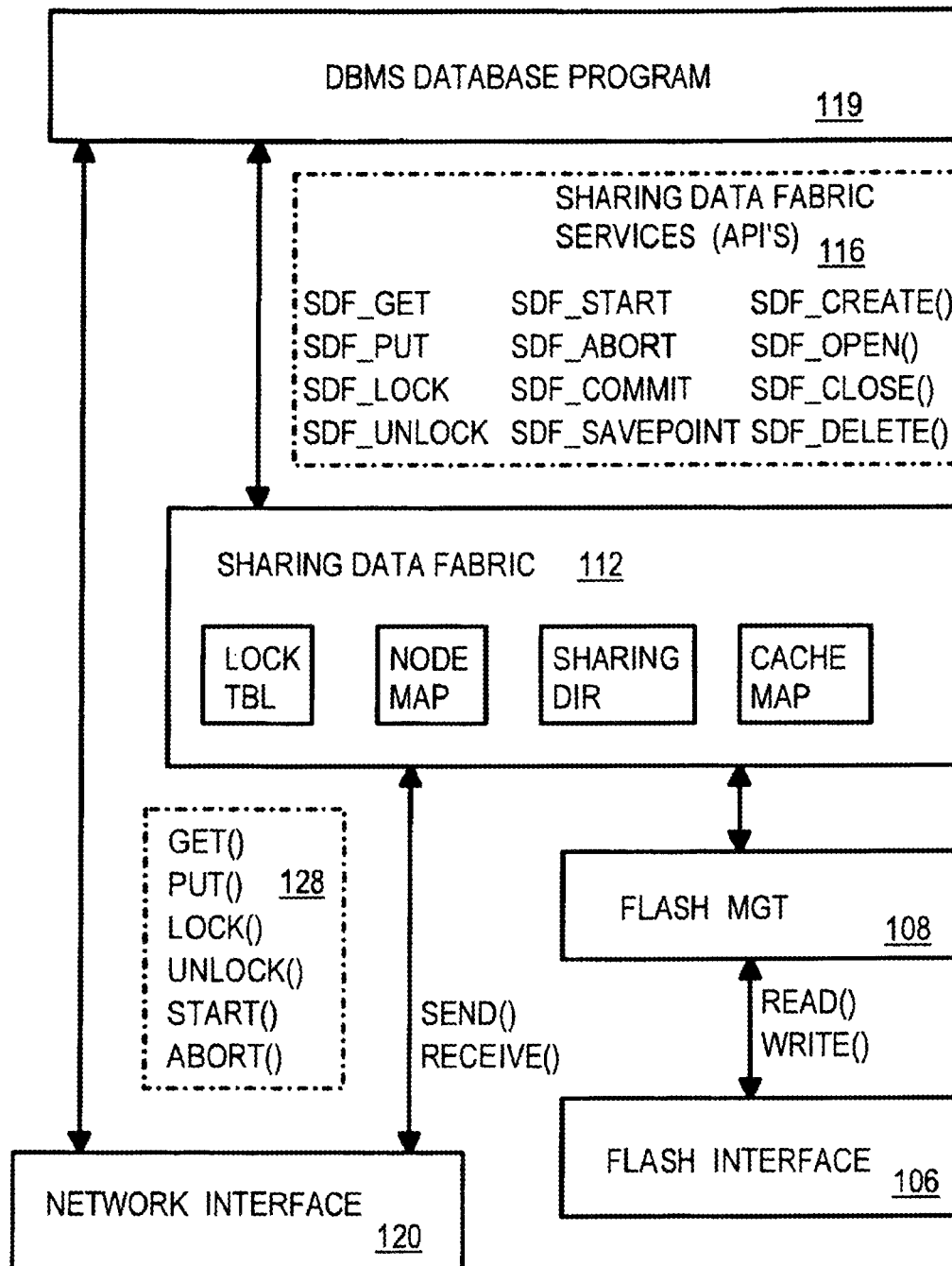


FIG. 8

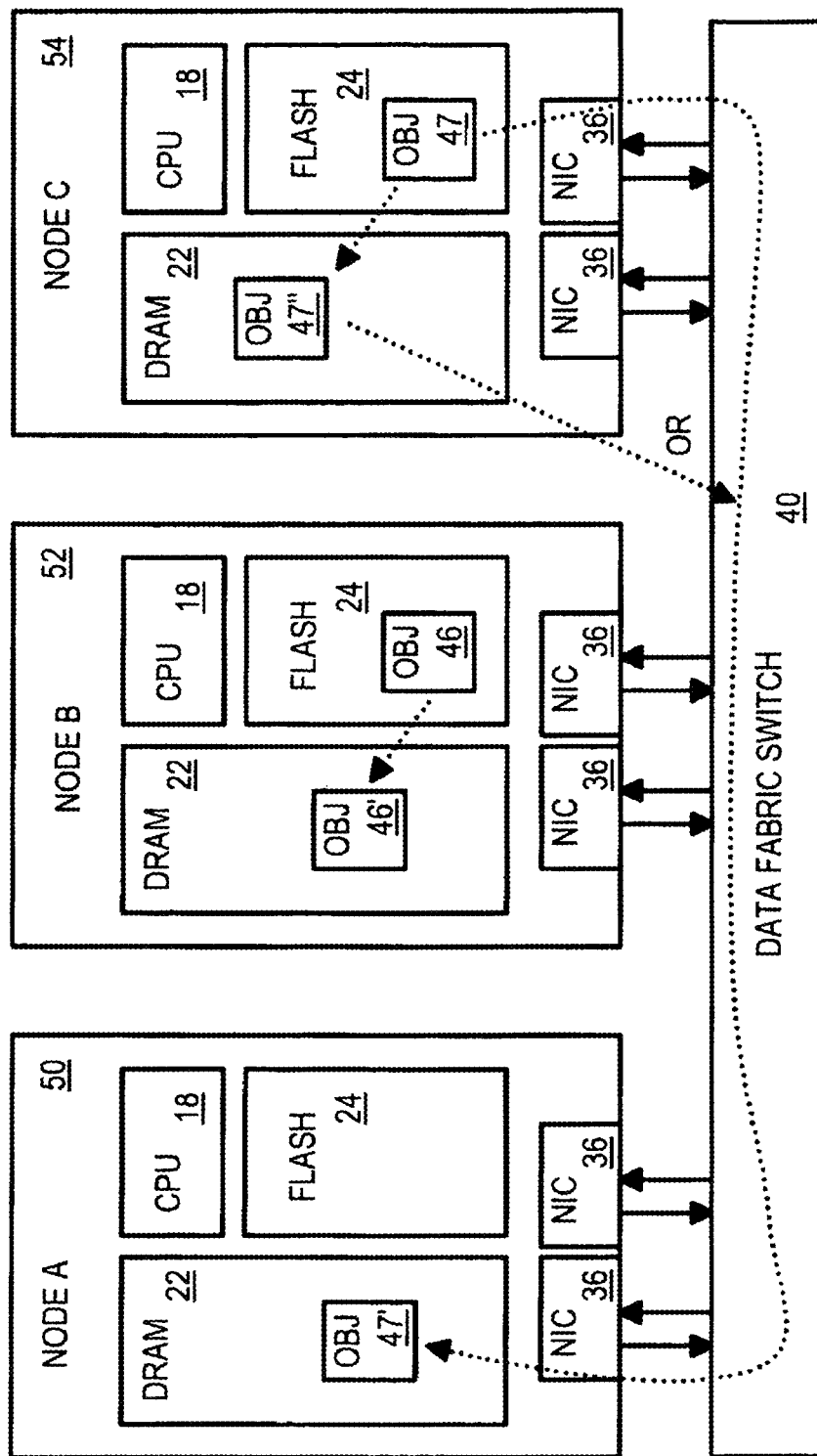


FIG. 9

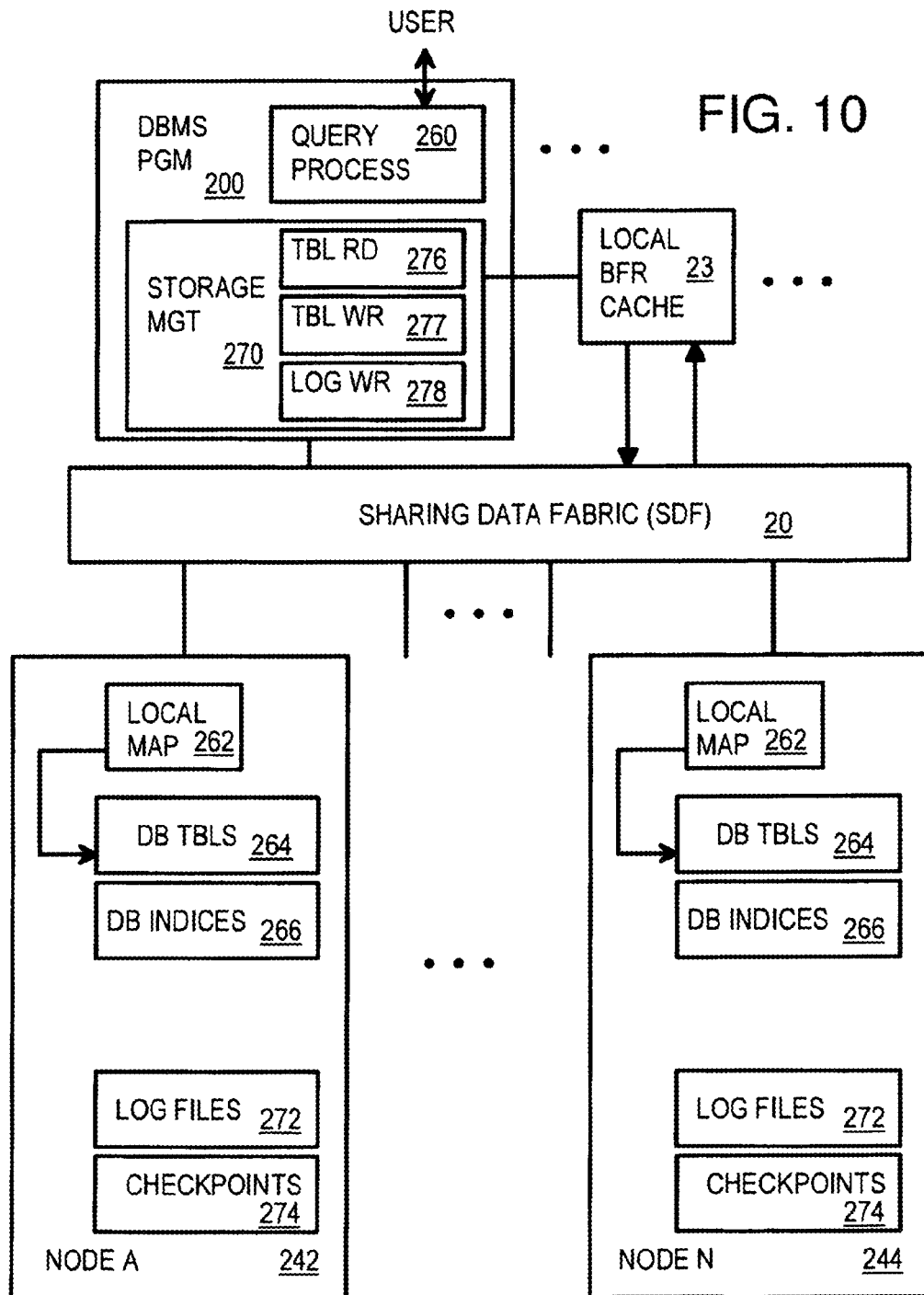
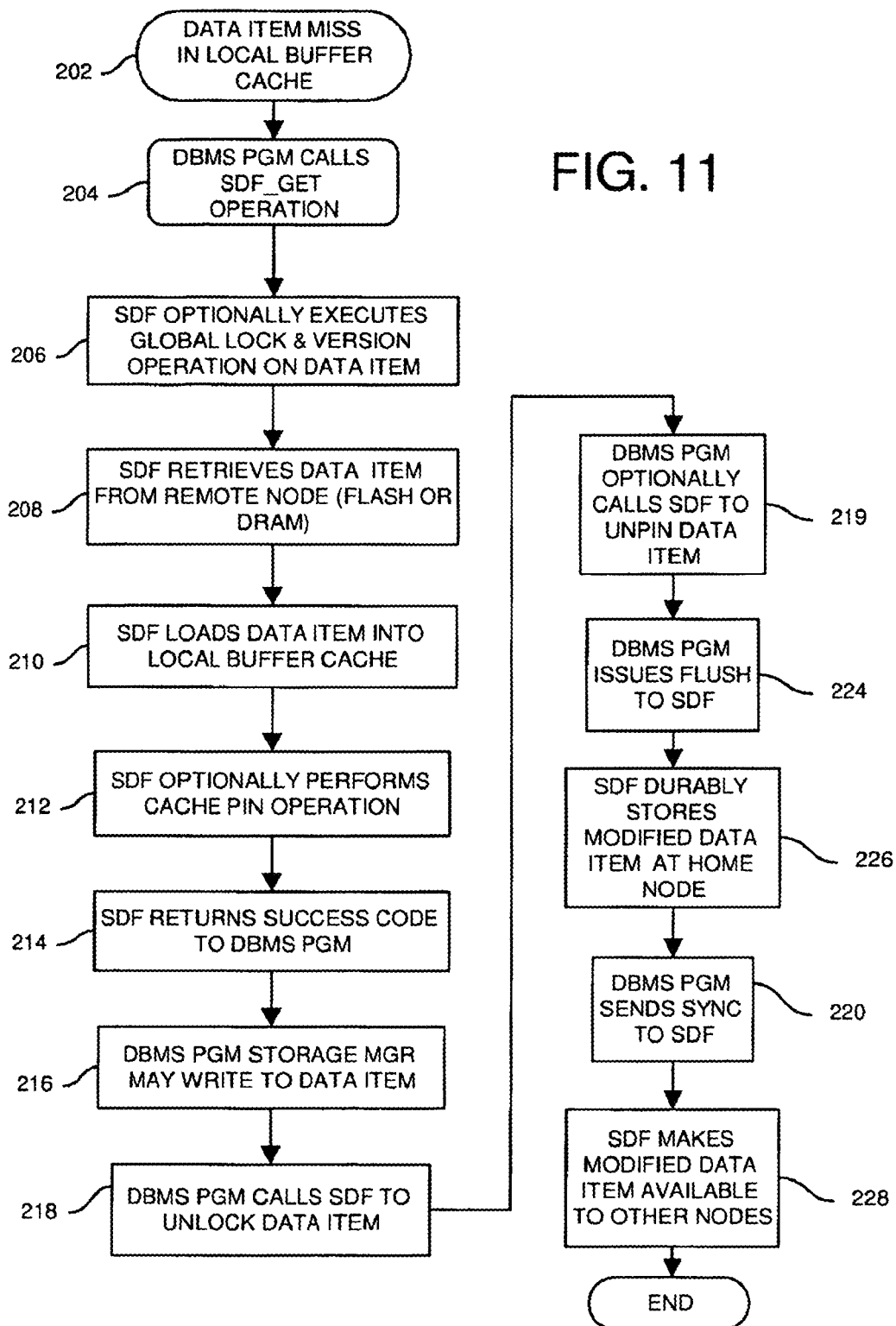


FIG. 11



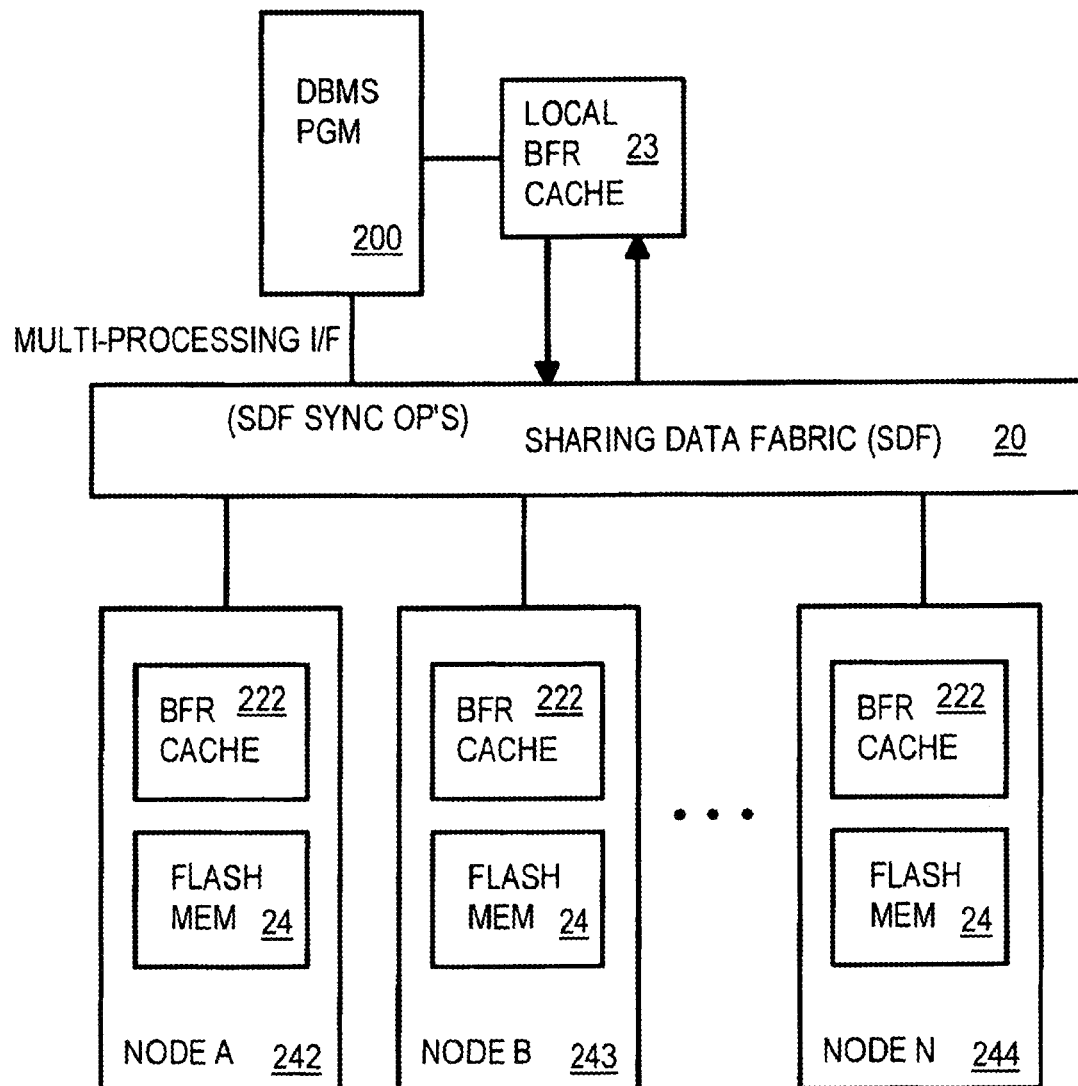


FIG. 12A

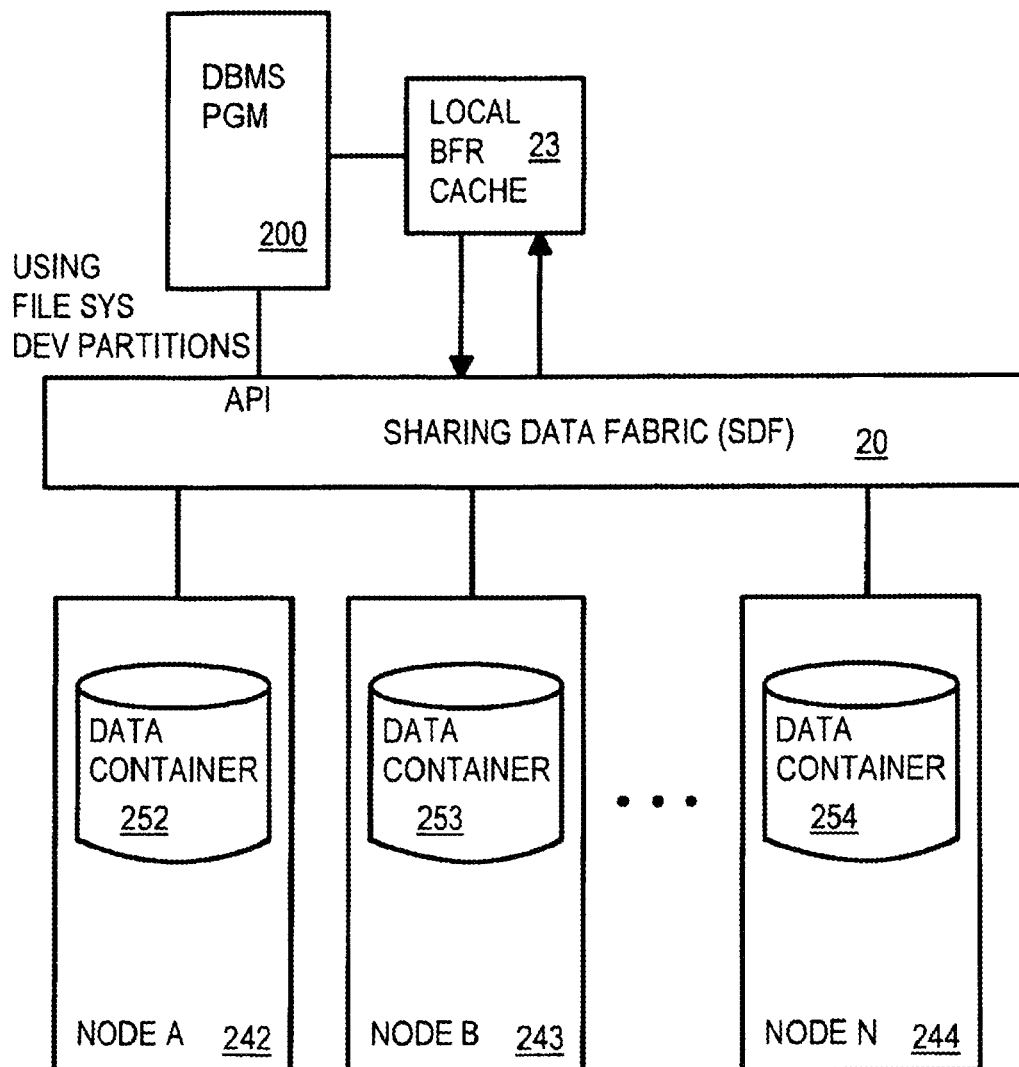


FIG. 12B

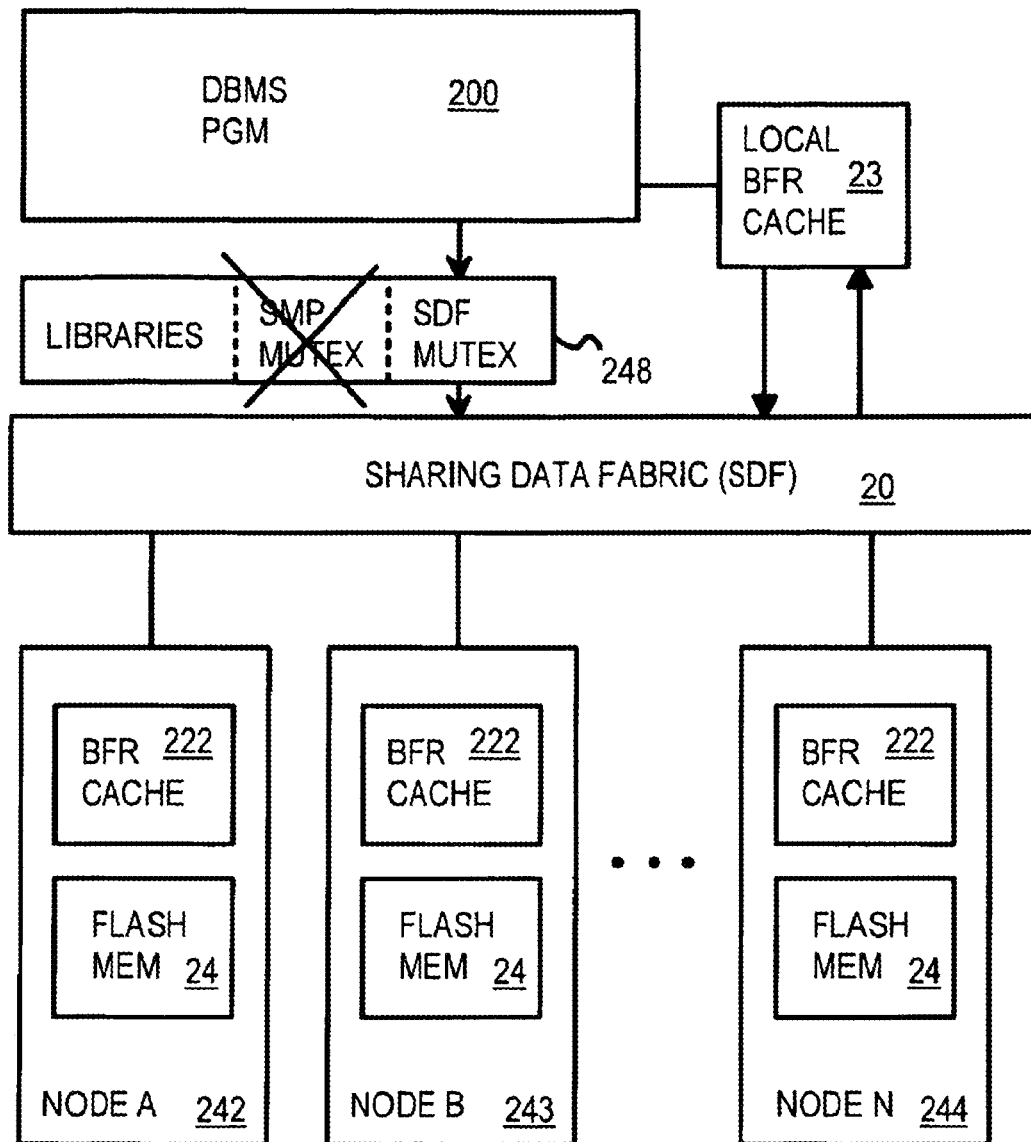


FIG. 12C

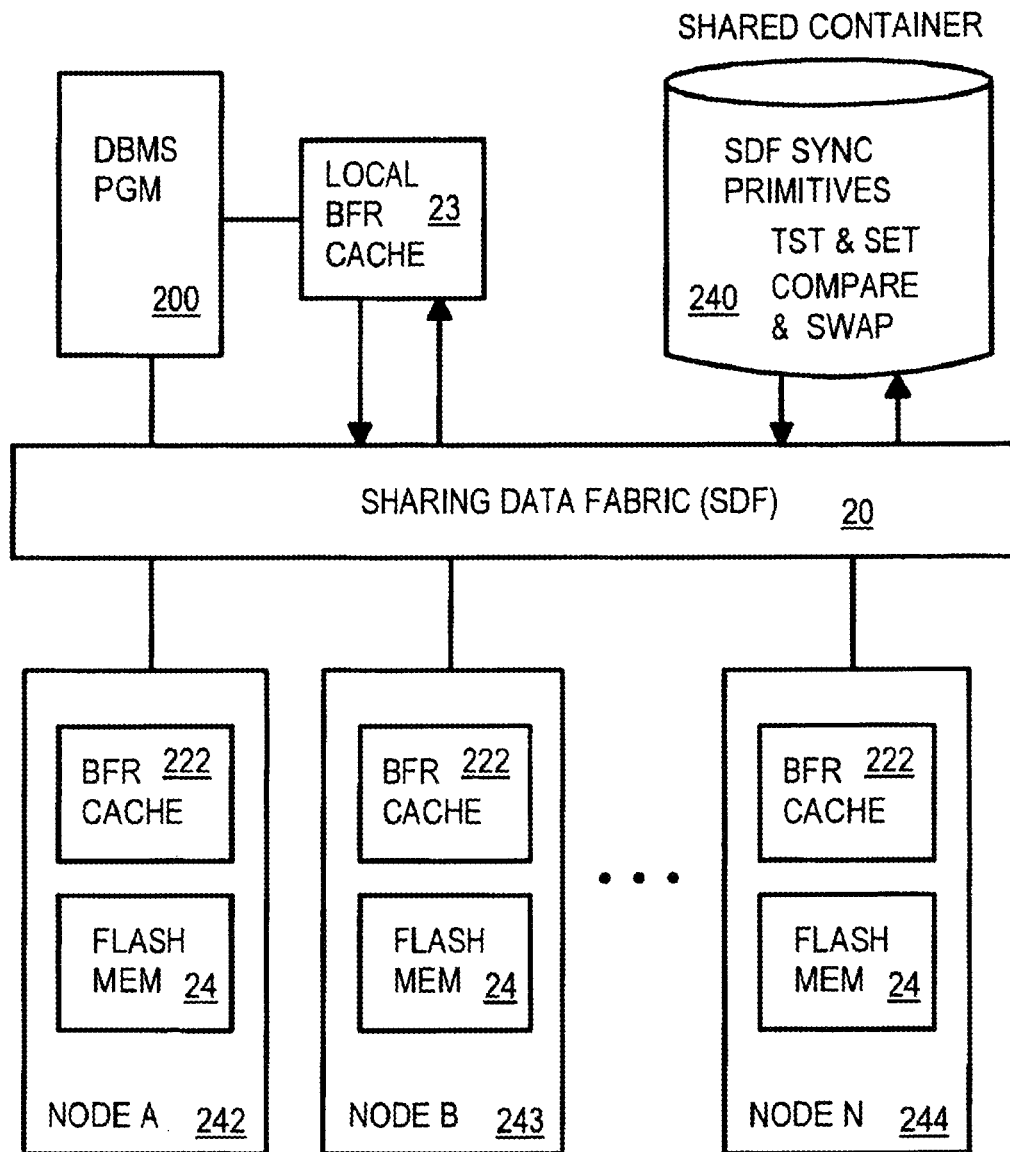


FIG. 12D

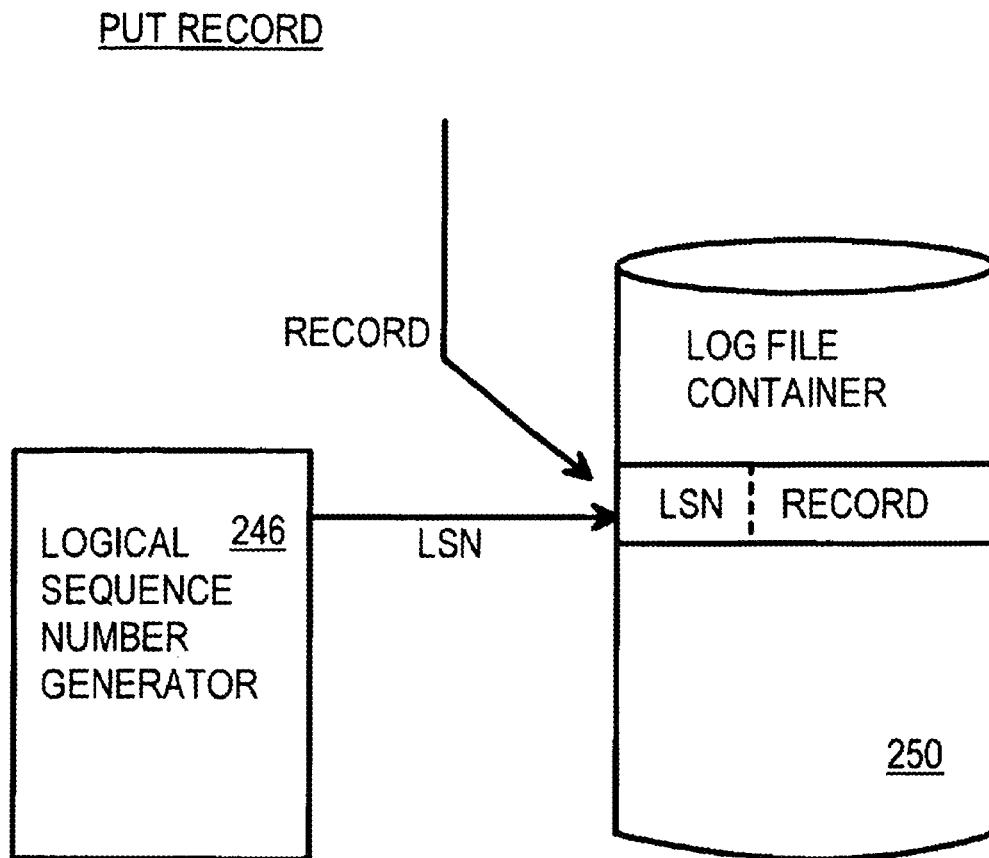


FIG. 13

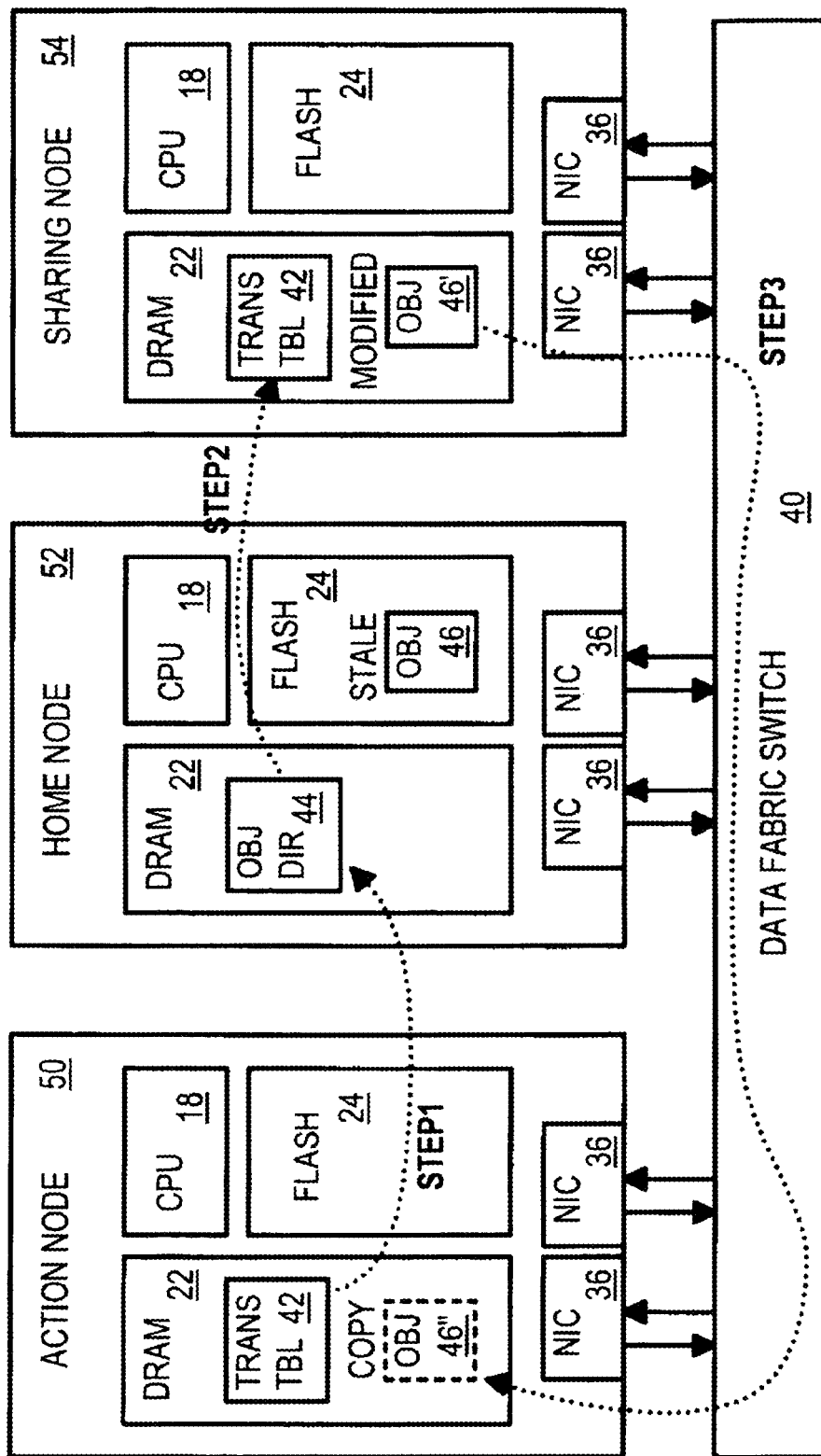


FIG. 14

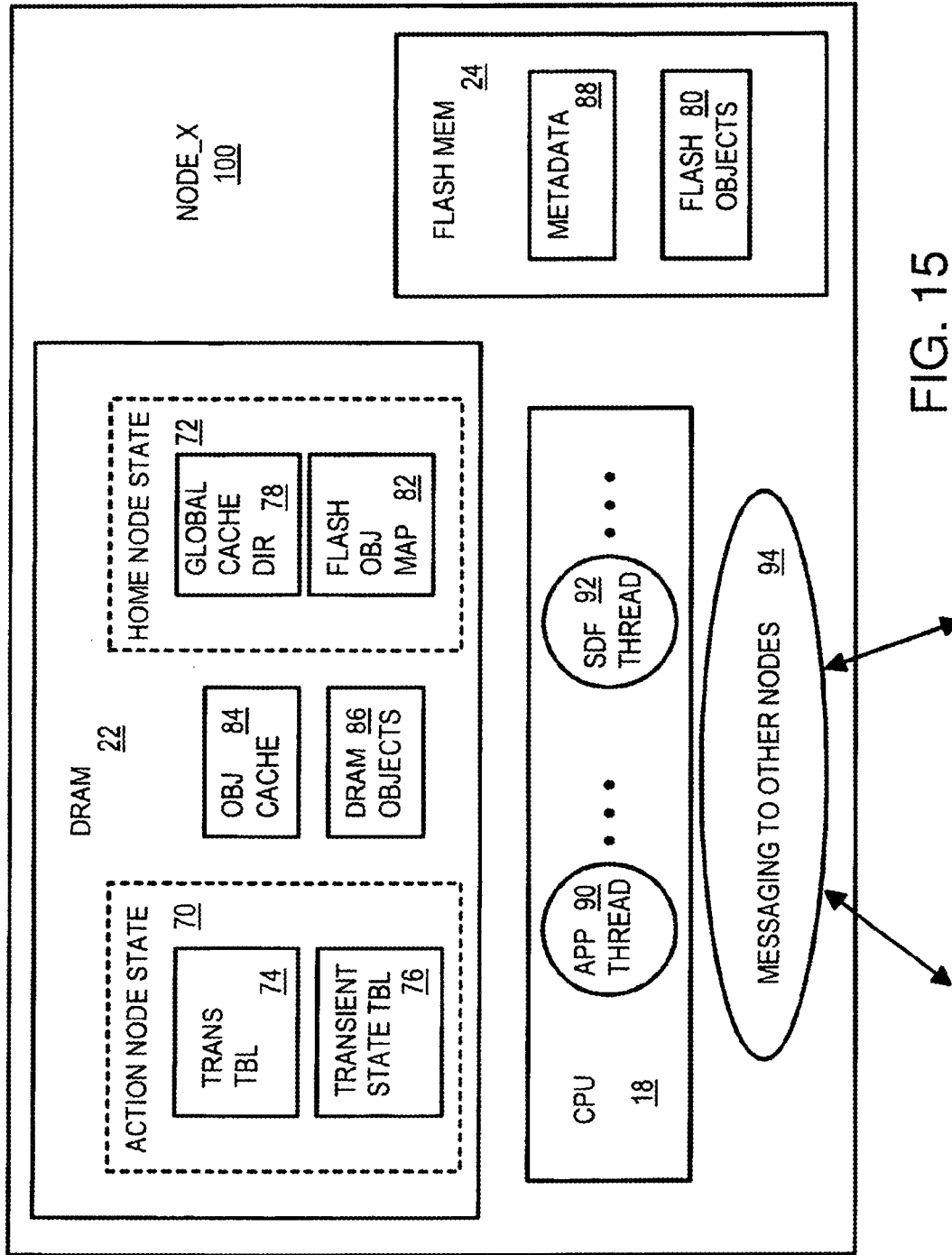
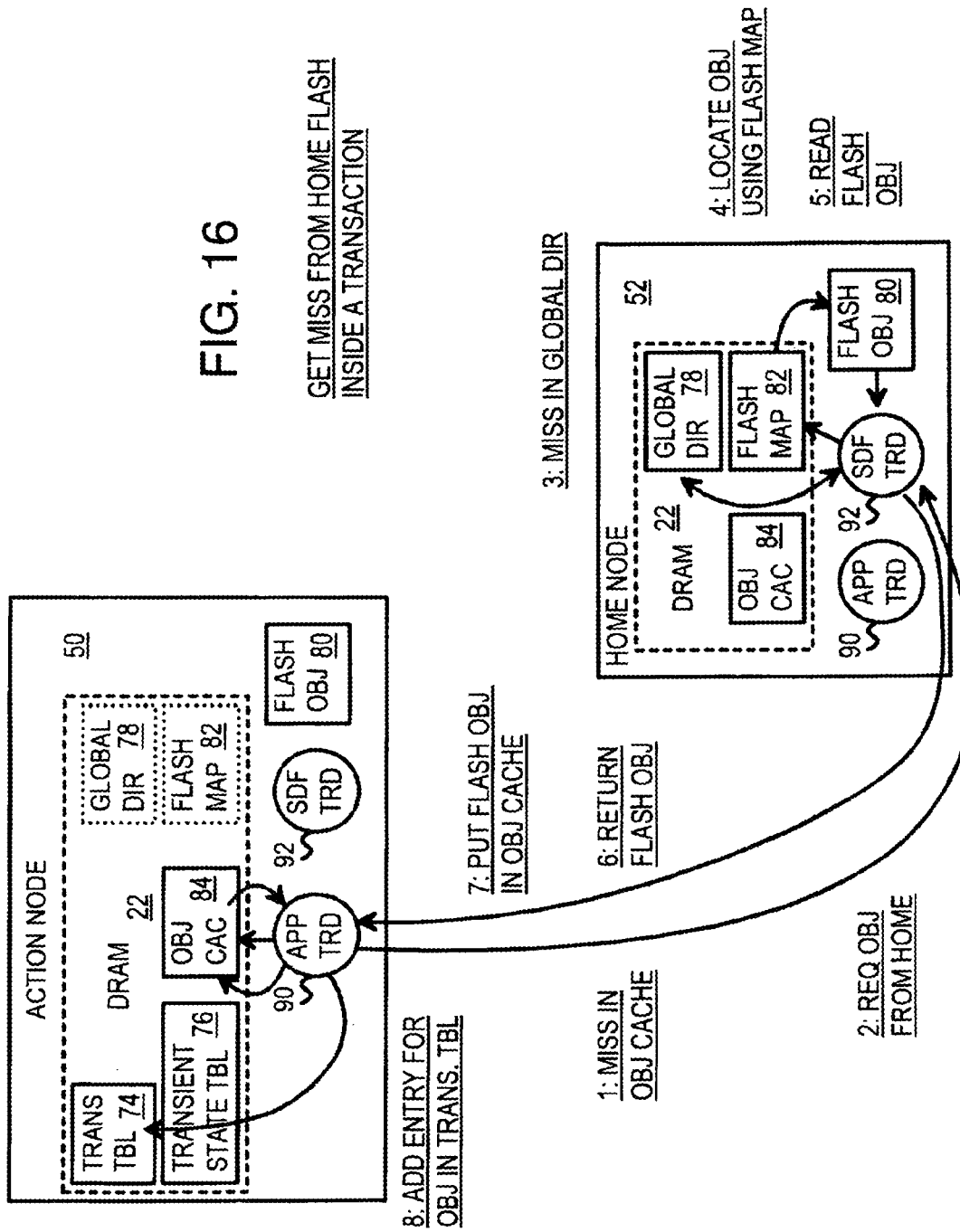
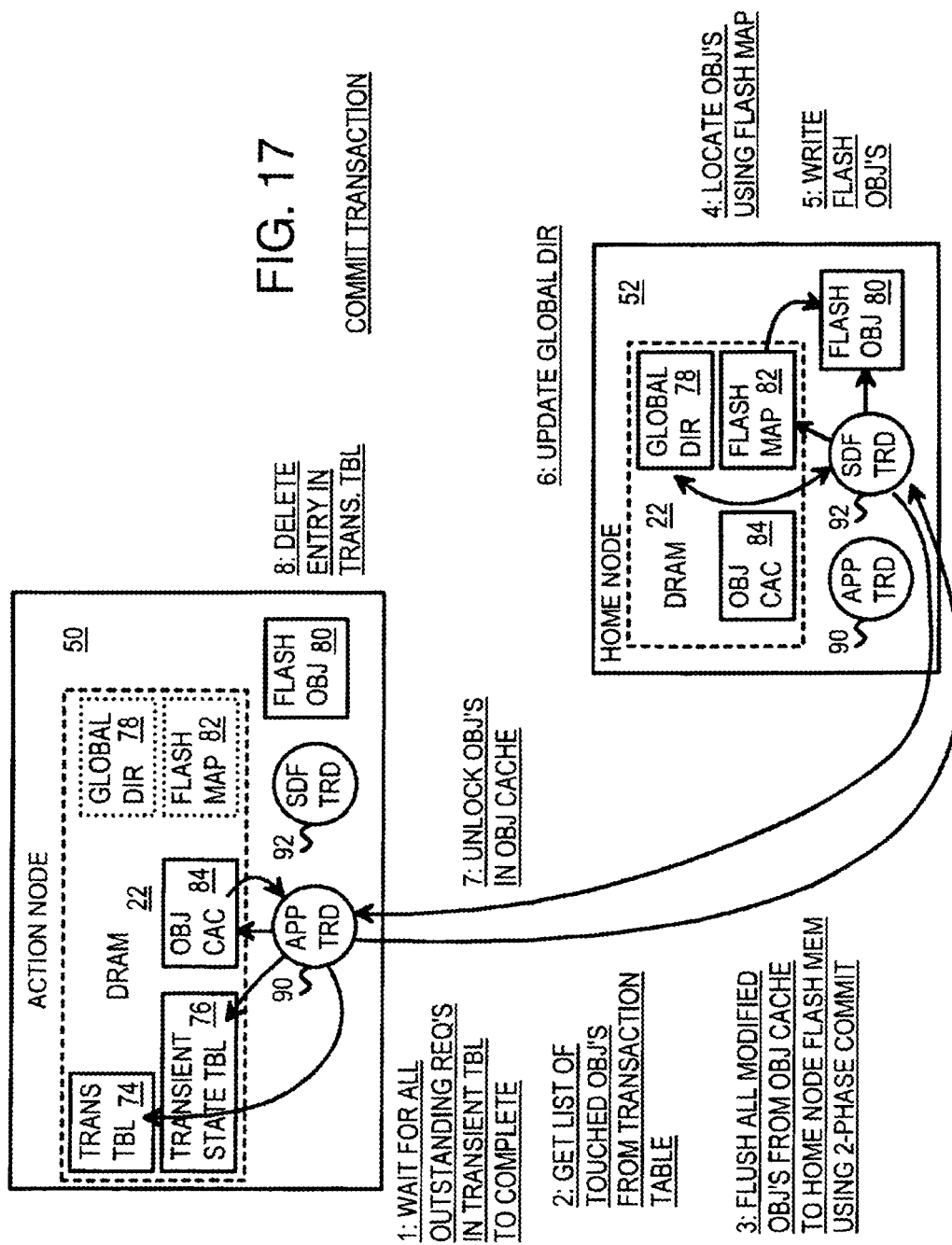
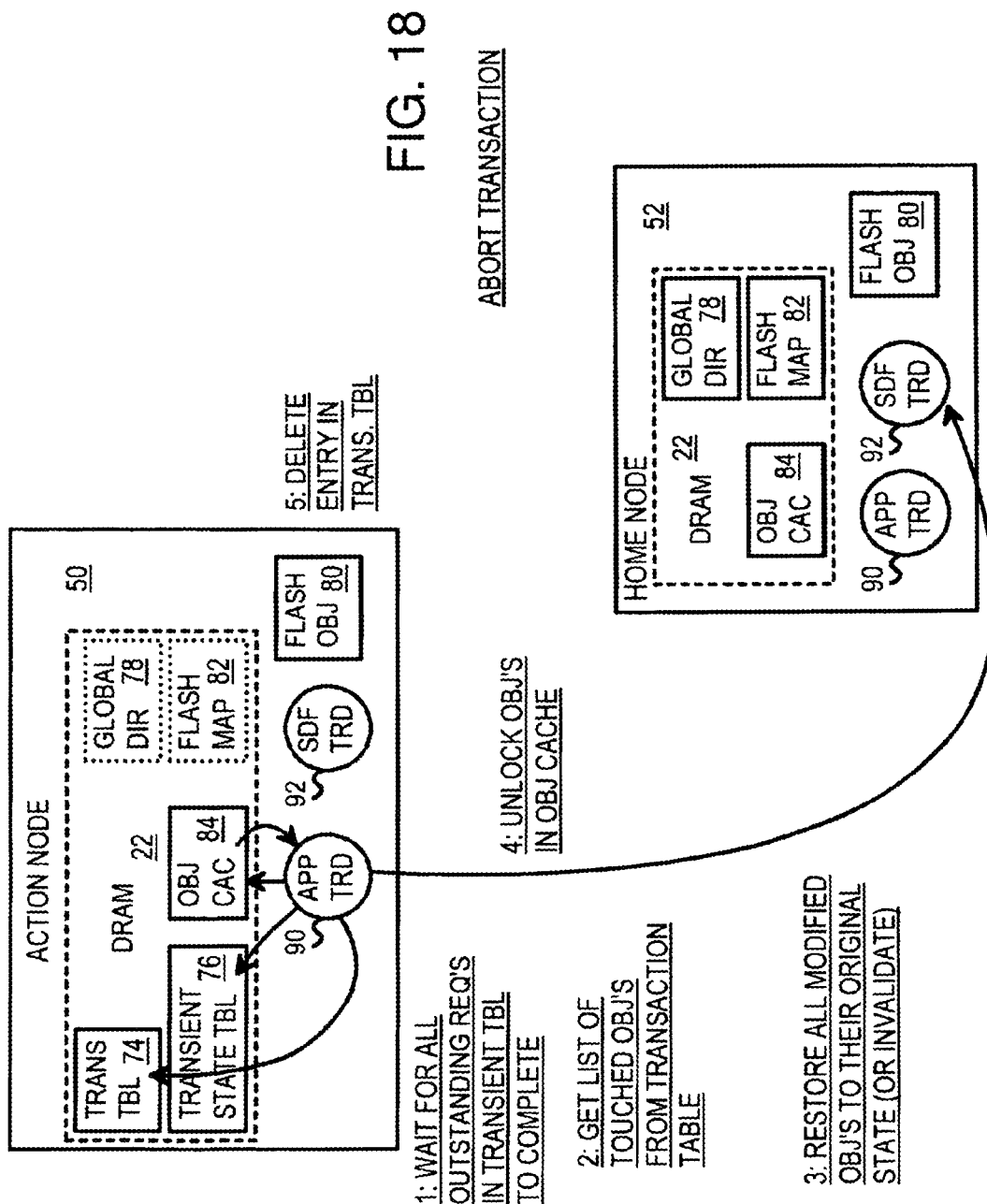


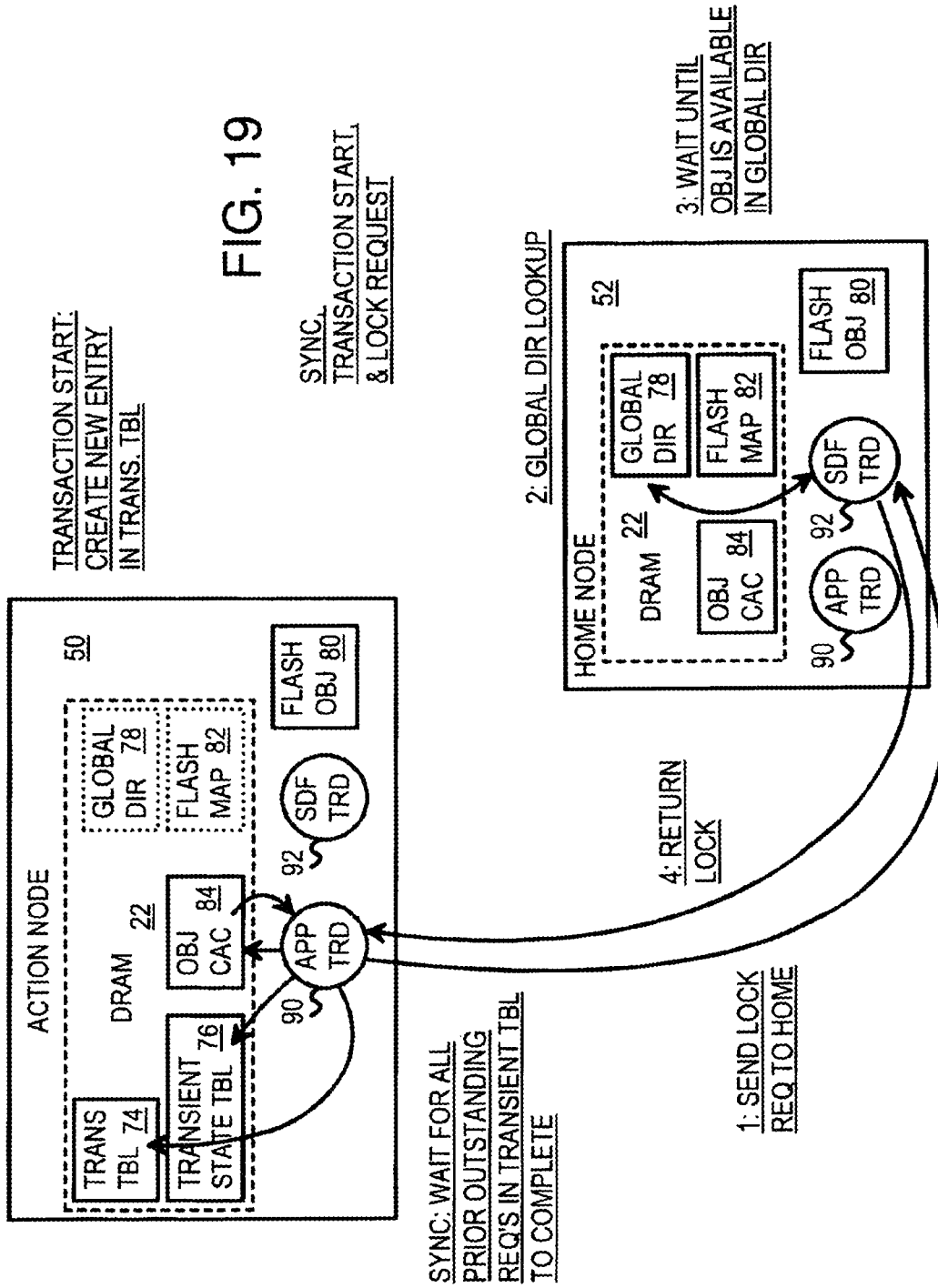
FIG. 15

FIG. 16









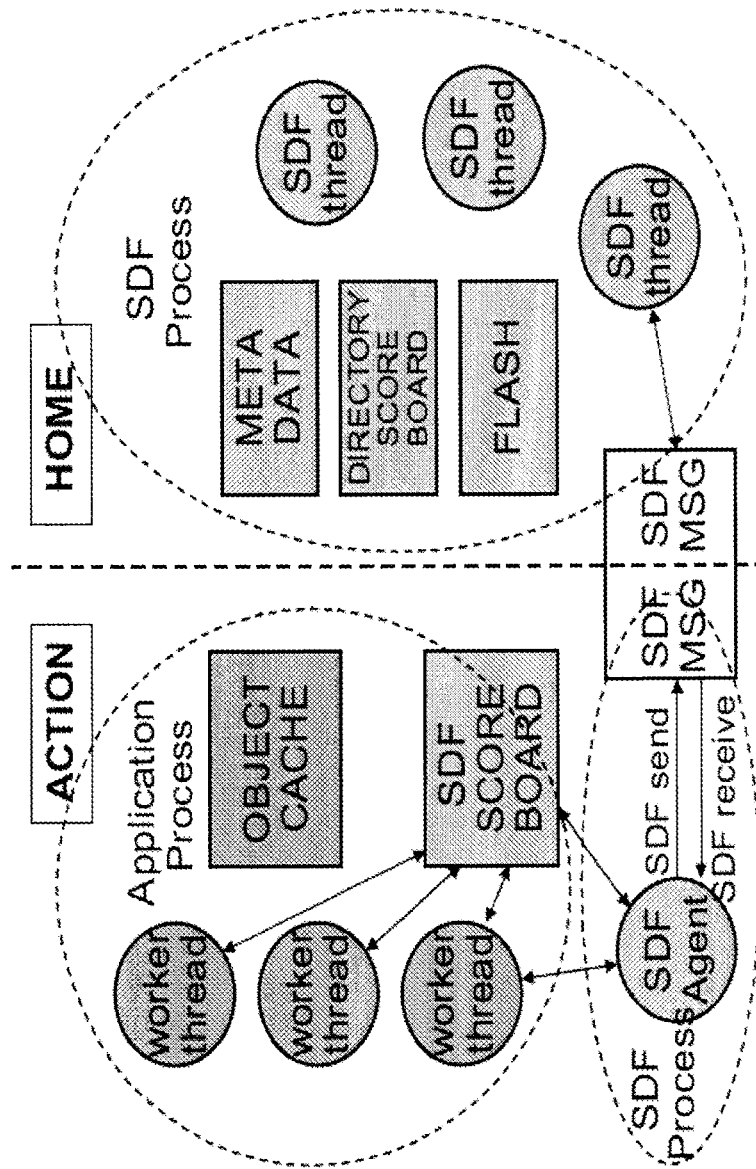


FIG. 20

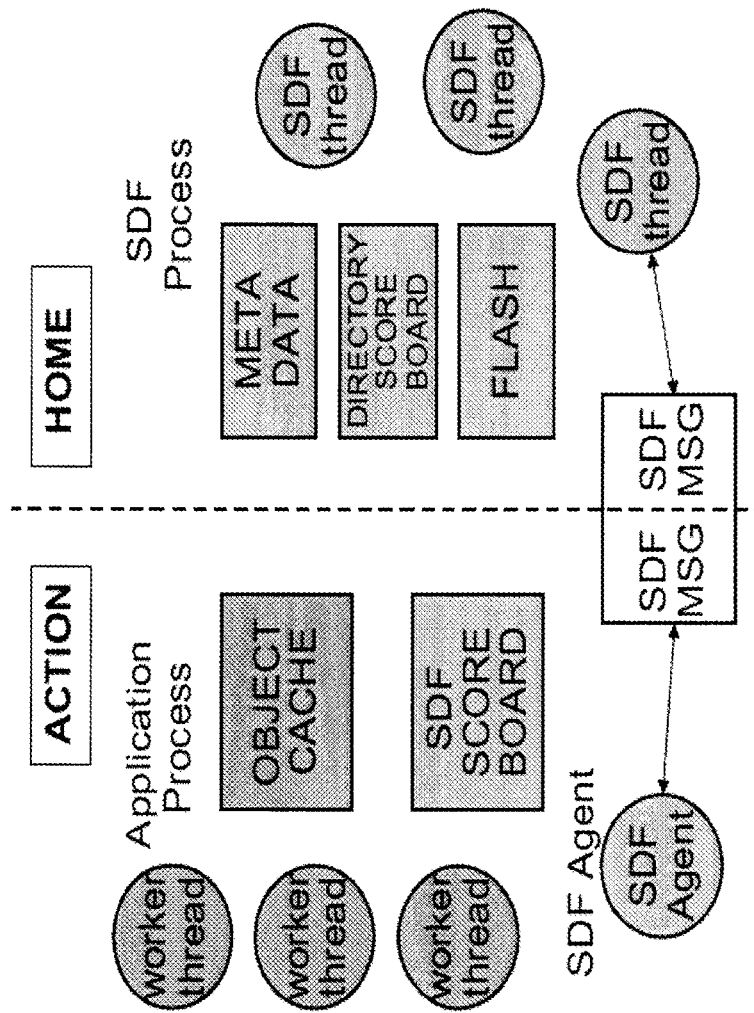


FIG. 21

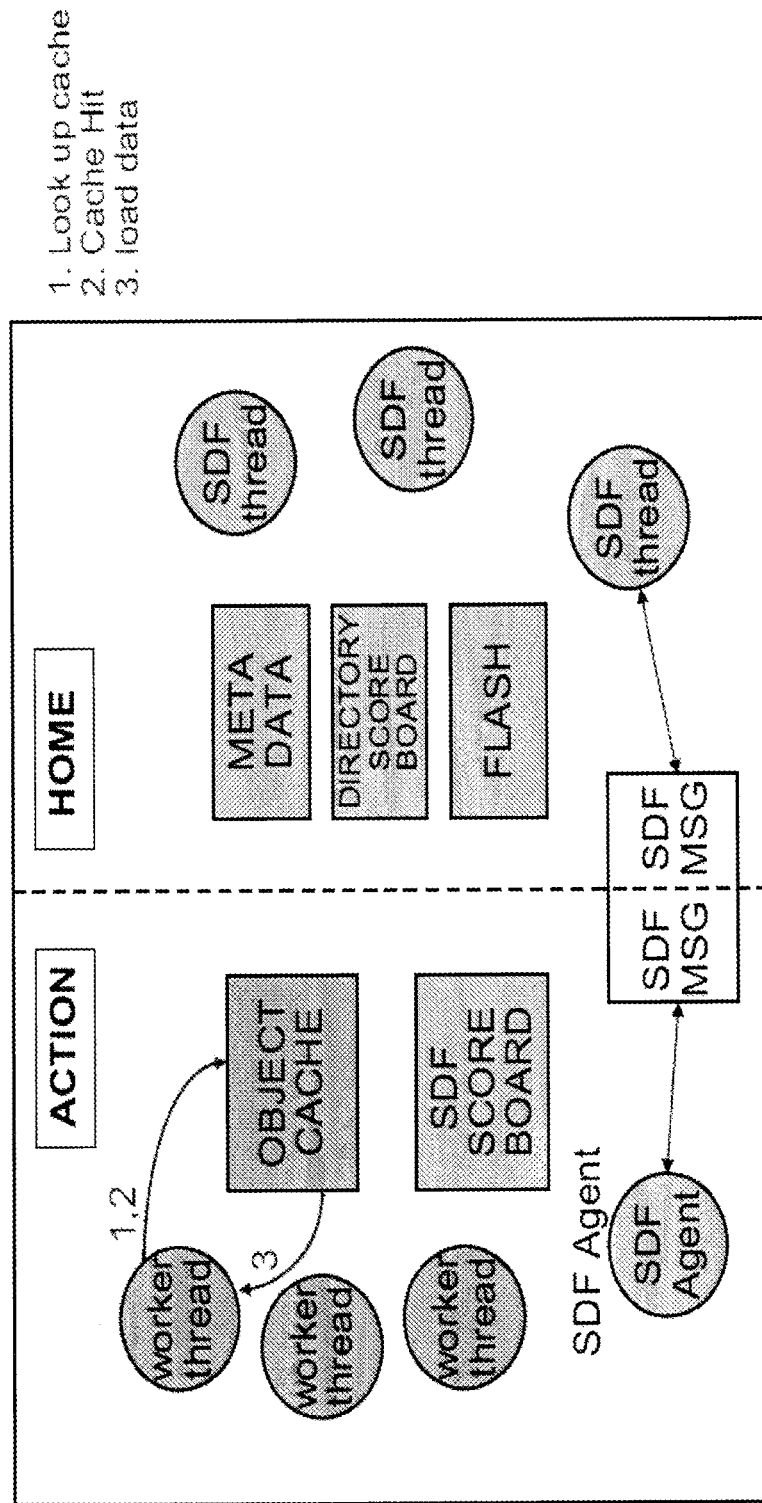


FIG. 22

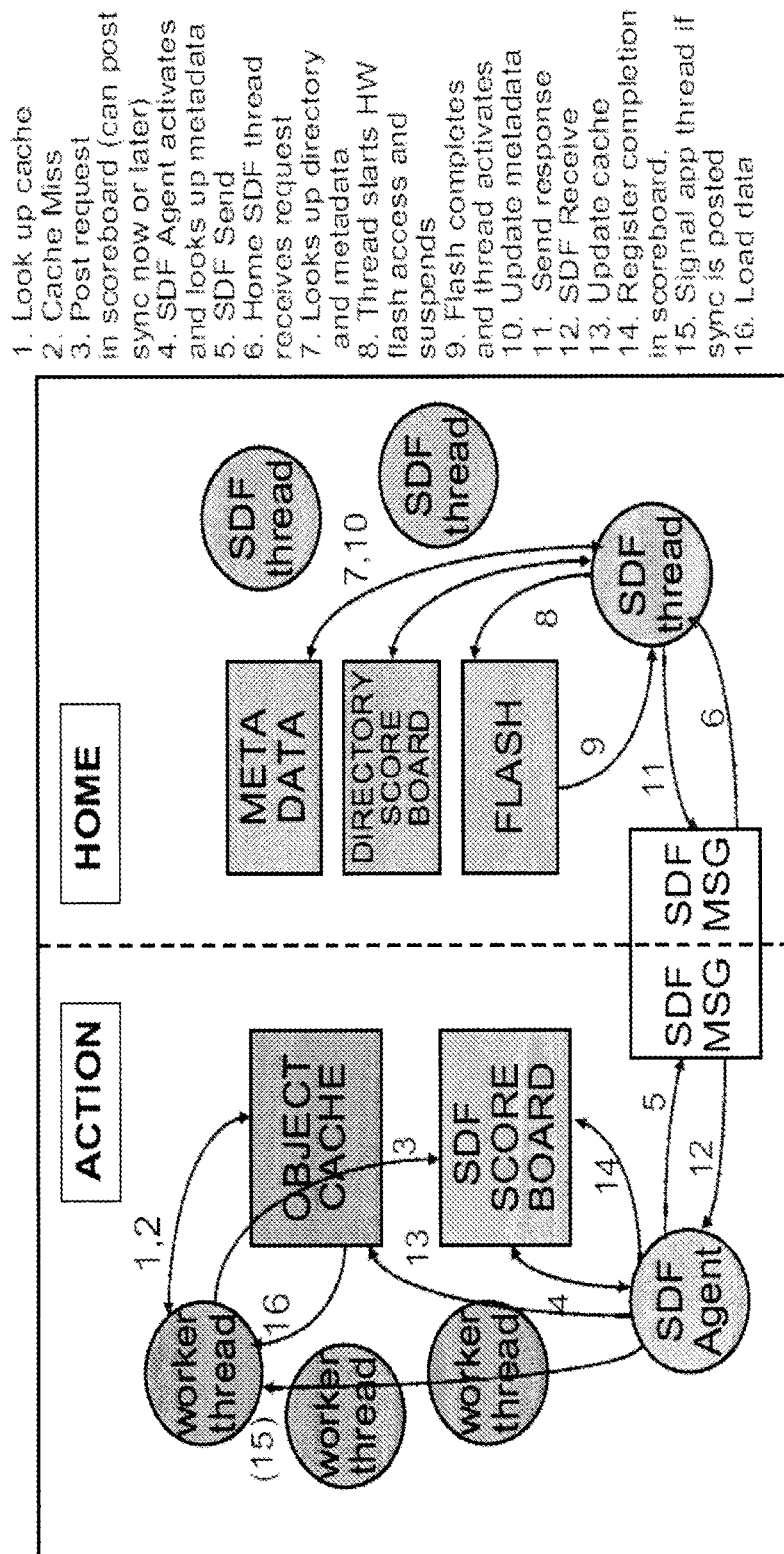


FIG. 23

OP Code (GetMiss, PutBack, Flush...)	Operation to be performed
SDF Address (Container, Object ID...)	SDF Container Address (+ shard, pnode)
Cache Index	Destination pointer for object data
Thread ID	Thread to signal on completion w/sync
Sync Flag	Flag for completion notification needed
Completion Status	Completion status; 0 = not complete

FIG. 24

User application SDF Library

req = **Post_Request**(OPcode, SDFaddr, Cacheindex, Sync)
- also { *req*=**newReq**; *req*.OPcode=GETMISS, etc; **Post**(*req*) }
- sends event notification to SDF Agent with *req* (request block index)
completion_status = **Sync_Wait**(*req*list) list of request blocks
- if completions occurred, return completion status vector
else posts Sync Flag and suspend on SDF agent notification
completion_status = **Sync_Poll**(*req*list)
- returns current completion status vector
Done(*req*list)
- Frees request blocks in the scoreboard; cleanup

Request Block in Scoreboard
Shared Memory

OP Code (GetMiss, PutBack, Flush...)
SDF Address (Container, Object ID...)
Cache Index
Thread ID
Sync Flag
Completion Status

SDF Agent

req = **Get_Request**(next)
- reads request block index from notification queue
- can read request fields using *req*.OPcode etc.
sync = **Post_Response**(*req*)
- posts the completion status in request block
if Sync Flag is set, send event to *req*.Thread ID

FIG. 25

User application SDF Library

req = Post_Request(OP=GetMiss,Container=FOO, pnode=zzyx,..., Cacheindex= freeblock(), Sync=NOW)

GetMiss	OP Code
FOO, zzyx, shard334, OBJ12345,+22	SDF Address
0xFF533214	Cache Index
3317	Thread ID
1	Sync Flag
0	Completion Status

FIG. 26

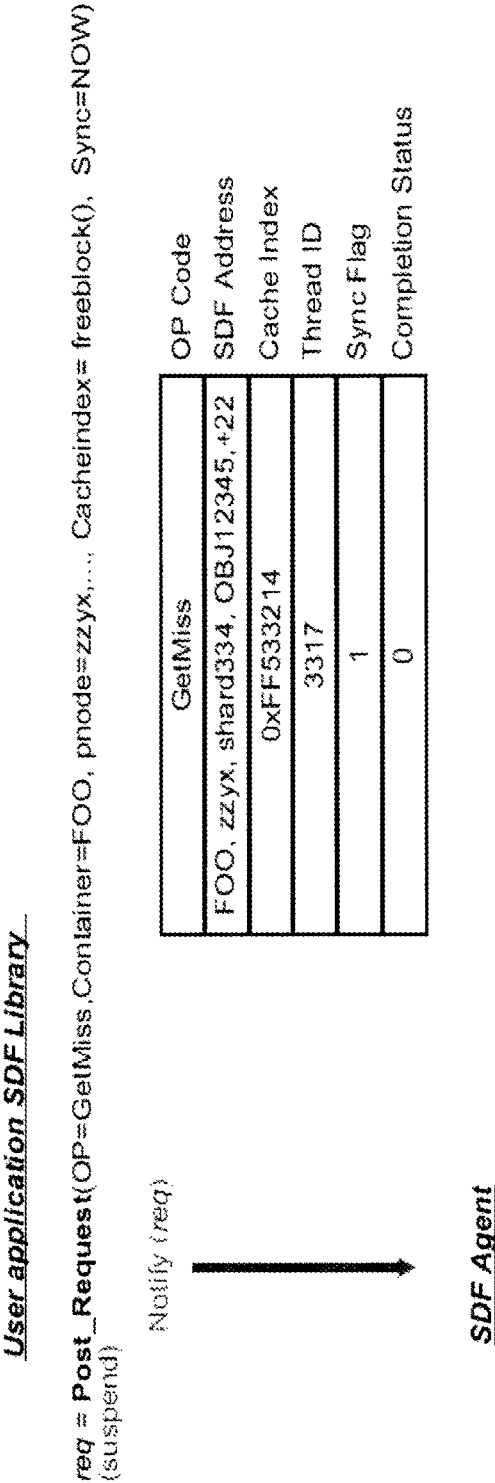


FIG. 27

User application SDF Library

GetMiss	OP Code
FOO, zzyx, shard334, OBJ12345,+22	SDF Address
0xFF533214	Cache Index
3317	Thread ID
1	Sync Flag
0	Completion Status

SDF Agent

```
req = Get_Request(next)
message.OPcode = req.OPcode
message.pnode = req.pnode
message.request_block = req
...
SDF_send(message)
```

FIG. 28

User application SDF Library

GetMiss	OP Code
FOO, zzyx, shard334, OBJ12345, +22	SDF Address
0xFF533214	Cache Index
3317	Thread ID
1	Sync Flag
1	Completion Status

SDF Agent

```
message = SDF_Receive(next)
req = message.request_block
bcopy(message.data, req.Cacheindex, message.data_size)
sync = Post_Response(req, message.response_code)
if sync then signal(app_process, req.ThreadID)
```

FIG. 29

User application SDF Library



Notify(req)

SDF Agent

```
message = SDF_Receive(next)
req = message.request_block
bcopy(message.data, req.Cacheindex, message.data_size)
sync = Post_Response(req, message.response_code)
if sync then signal(app_process, req.threadID)
```

OP Code	GetMiss
SDF Address	FOO, zzyx, shard334, OBJ12345, +22
Cache Index	0xFF533214
Thread ID	3317
Sync Flag	1
Completion Status	1

FIG. 30

User application SDF Library

```
req = Post_Request(OP=GetMiss,Container=FOO, pnode=zzyx,..., CacheIndex= freeblock(), Sync=NOW)
(resume)
if req.CompletionStatus == OK then return(req.CacheIndex)
```

GetMiss	OP Code
FOO, zzyx, shard334, OBJ12345,+22	SDF Address
0xFF533214	Cache Index
3317	Thread ID
1	Sync Flag
1	Completion Status

FIG. 31

Examples of such synchronization and ordering constructs are:

- mem_ref_t *lock_alloc()
void lock(mem_ref_t *l)
void lock_dealloc(mem_ref_t *l)
- fetch-and-op(mem_ref_t *p, op, arg, size)
Get the data of length "size" at shared memory location p, and then perform the specified operation "op" on the data at location "p" using the additional data provided by arg:
For example, "fetch-and-op(p, ADD, 1, 4)" returns the 4-byte value located in shared memory at location "p", then adds 1 to the value in shared memory.
This form of synchronization primitive is commonly used to implement efficient critical sections, statistics counters, and linked lists.
- sync(mem_ref_t *p)
Ensure that all prior accesses to shared memory location p have been completed. This is useful if the shared memory implementation exploits the higher performance that is possible in weakly ordered hardware systems.

FIG. 32

- Creation/Deletion/Attachment of Shared Memory Arenas:

```
/**
 * Create memory allocation arena arena_name using a single mmap'd
 * file for backing store. This has the potential to allow large
 * object allocations to occupy non-contiguous pages but requires
 * a ramdisk or tmpfs backing store to avoid lazy page writes to
 * disk.
 *
 * Returns
 * 0 on success
 * MEM_ARENA_EEXIST on failure
 */
int mem_arena_create_mmap_file(
    const char *arena_name,
    size_t size_limit,
    const char *backing_file);

/**
 * Remove the named memory allocation arena from the system
 */
int mem_arena_remove(const char *arena_name);

/**
 * Attach to a specific memory allocation arena with namespace
 * shared across the node.
 *
 * mem_arena_detach() should be called on termination for diagnostic
 * purposes but for correctness both abnormal and normal termination
 * imply detach.
 *
 * Returns
 * MEM_ARENA_REF_ENOENT on failure
 */
mem_arena_ref_t mem_arena_attach(const char *arena_name);

/**
 * Detach from arena. All mem_ref_t references associated with
 * arena_ref are released.
 */
void mem_arena_detach(mem_arena_ref_t arena_ref);
```

FIG. 33

 - Allocate/Release a Piece of Memory from an Arena:

```

/**
 * Allocate an object with one reference attached to
 * arena_ref, returning that reference. policy
 * specifies retention/protection poli
 *
 * PROT_ANY - permissions can be manipulated arbitrarily
 * PROT_WRITE_ONCE - the object may be set to read-only and
 * stay there
 * PROT_LRU - retain when reference count hits zero. LRU
 * eviction is explicitly specified so dangling weak
 * references manifest as bugs more often.
 *
 * Memory allocation/referencing and mapping are separate
 * to accomodate proxy operation, early failures on resource
 * shortage, and applications with 32 bit address spaces.
 *
 * Returns
 * MEM_REF_ENOMEM on failure
 */
mem_ref_t mem_alloc(mem_arena_ref_t arena_ref, size_t size,
    int prot_policy,
    int flags);
/**
 * Release a memory reference. The objects will be lazily
 * released from memory allowing use as a cache.
 */
void mem_ref_release(mem_ref_t);
/**
 * Guarantee a map of the range [offset, offset + length) exists
 * within this context's address space. prot is per mmap(2)
 *
 * Calls to mem_ref_map may be idempotent.
 *
 * Returns
 * Mapped address on success
 * MEM_VOID_PTR_EACCESS on protection failure
 */
void *mem_ref_map(mem_ref_t ref, size_t length, int prot,
    off_t offset);
/**
 * Suggest a given reference mapping is unnecessary although a
 * physical unmapping would aid debugging
 *
 * XXX physical unmapping is only possible with mmap'd ramdisk
 * and not SYSV shm backing store.
 *
 * Returns:
 * 0 on success
 */
int mem_ref_unmap(mem_ref_t, void *addr);
/**
 * Set mem_ref access protections using PROT_ from sys/mman.h.
 * PROT_READ readable
 * PROT_WRITE writeable
 * PROT_READ|PROT_WRITE read-write
 * PROT_NONE no permissions
 *
 * Permissions may not be checked until the next mem_ref_clone().
 *
 * Returns:
 * 0 on success
 * MEM_EACCESS on an illegal permission.
 */
int mem_ref_set_prot(mem_ref_t ref, int prot);

```

FIG. 34

- Interprocess Identifiers and Multiversioning of Objects

```
/**
 * Return the unique ID for ref which can be passed between contexts.
 * Different versions may be assigned to a given permanent id.
 * Reference count hitting zero causes mem_id_ref() to fail.
 */
mem_id_t mem_ref_get_perm_id(mem_ref_t ref);

/**
 * Return a unique ID for the current version of ref which can be
 * passed between contexts. The reference count hitting zero causes
 * mem_id_ref() to fail.
 */
mem_id_t mem_ref_get_current_id(mem_ref_t ref);

/**
 * Create a local reference from id. MEM_ID_ROOT refers to the
 * permanent ID of the root object for the pool.
 *
 * Returns:
 *   MEM_REF_ENOENT if the reference count has hit zero before this
 *   call.
 *   MEM_REF_EACCESS if the object's permissions are incorrect
 */
mem_ref_t mem_id_ref(mem_id_t id, int prot);

/**
 * Change the data version associated with a permanent id. Used
 * to complete updates when using multi-version concurrency control.
 * Does not affect the permanent id associated with ref.
 *
 * Returns:
 *   0 on success
 *   MEM_ENOENT when perm_id does not exist
 *   MEM_EVERSION where the current id associated with perm_id
 *   does not match.
 */
int mem_ref_set_current_id(mem_ref_t ref, mem_id_t perm_id,
                           mem_id_t equals_current_id);
```

FIG. 35

- Copy Shared Memory Objects

```
/**
 * Duplicate a reference so that different subsystems within one
 * address space can reference count objects. This is safer
 * than incrementing and decrementing reference counts.
 */
mem_ref_t mem_ref_shallow_copy(mem_ref_t);

/**
 * Provide semantics equivalent to duplicating the object referred
 * to by a reference (eventually we may use the VM subsystem for
 * COW) to allow making a writeable copy of a read-only object.
 * policy specifies retention/protection policies
 *
 * PROT_ANY - permissions can be manipulated arbitrarily
 * PROT_WRITE_ONCE - the object may be set to read-only and
 * stay there
 * PROT_LRU - retain when reference count hits zero. LRU
 * eviction is explicitly specified so dangling weak
 * references manifest as bugs more often.
 *
 * Returns:
 * MEM_REF_ENOMEM on allocation failure
 * MEM_REF_EACCESS when copy is not permitted XXX is this
 * reasonable?
 */
mem_ref_t mem_ref_deep_copy(mem_ref_t ref, int policy);
```

FIG. 36

1

**CLUSTER OF PROCESSING NODES WITH
DISTRIBUTED GLOBAL FLASH MEMORY
USING COMMODITY SERVER
TECHNOLOGY**

**CLAIM OF PRIORITY AND RELATED
APPLICATION DATA**

The present application claims priority to U.S. provisional patent application 61/323,302, entitled "Cluster of Processing Nodes with Distributed global Flash Memory Using Commodity Server Technology," filed on Apr. 12, 2010, and is hereby incorporated by reference for all purposes as if fully set forth herein.

This application also claims priority to U.S. provisional patent application No. 61/323,351, entitled "Distributed Data Access Using Solid State Storage," filed Apr. 12, 2010, invented by John Busch et al., the entire contents of which are incorporated by reference for all purposes as if fully set forth herein.

This application is related to U.S. non-provisional patent application Ser. No. 12/276,540, entitled "Scalable Database Management Software on a Cluster of Nodes Using a Shared-Distributed Flash Memory," filed on Nov. 24, 2008, invented by Darpan Dinker et al., the entire contents of which are incorporated by reference for all purposes as if fully set forth herein.

This application is related to U.S. non-provisional patent application Ser. No. 12/983,754, entitled "Efficient Flash Memory-Based Object Store," filed on Jan. 3, 2011, invented by John Busch et al., the entire contents of which are incorporated by reference for all purposes as if fully set forth herein.

This application is related to U.S. non-provisional patent application Ser. No. 12/983,758, entitled "Flexible Way of Specifying Storage Attributes in a Flash-Memory Based Object Store," filed on Jan. 3, 2011, invented by Darryl Ouye et al., the entire contents of which are incorporated by reference for all purposes as if fully set forth herein.

This application is related to U.S. Non-provisional patent application Ser. No. 12/983,762, entitled "Minimizing Write Operations to a Flash Memory-Based Object Store," filed on Jan. 3, 2011, invented by Darpan Dinker, the entire contents of which are incorporated by reference for all purposes as if fully set forth herein.

This application is related to U.S. non-provisional patent application Ser. No. 13/084,368, now U.S. Pat. No. 8,868,487, entitled "Event Processing in a Flash Memory Based Object Store," filed Apr. 11, 2011, invented by Manavalan Krishnan et al., the entire contents of which are incorporated by reference for all purposes as if fully set forth herein.

This application is related to U.S. non-provisional patent application Ser. No. 13/084,432, entitled "Write Operations in a Flash Memory-Based Object Store," filed Apr. 11, 2011, invented by Xiaonan Ma et al., the entire contents of which are incorporated by reference for all purposes as if fully set forth herein.

This application is related to U.S. non-provisional patent application Ser. No. 13/084,511, now U.S. Pat. No. 8,793,531, entitled "Recovery and Replication of a Flash Memory-Based Object Store," filed Apr. 11, 2011, invented by Johann

2

George et al., the entire contents of which are incorporated by reference for all purposes as if fully set forth herein.

FIELD OF THE INVENTION

This invention relates to distributed object stores, and more particularly to, shared global memory accessible to object stores executing on a plurality of nodes using flash memory.

BACKGROUND OF THE INVENTION

Database programs are one of the most widely used and useful applications of computers. Data records may be stored in database tables that are linked to one another in a relational database. Queries from users allow database programs to locate matching records and display them to users for modification. Often a large number of users access different records in a database simultaneously.

Database records are typically stored on rotating hard disks. Computer hard-disk technology and the resulting storage densities have grown rapidly. Despite a substantial increase in storage requirements, disk-drive storage densities have been able to keep up. Disk performance, however, has not been able to keep up. Access time and rotational speed of disks, key performance parameters in database applications, have only improved incrementally in the last 10 years.

Web sites on the Internet may link to vast amounts of data in a database, and large web server farms may host many web sites. Storage Area Networks (SANs) are widely used as a centralized data store. Another widespread storage technology is Network Attached Storage (NAS). These disk-based technologies are now widely deployed but consume substantial amounts of power and can become a central-resource bottleneck. The recent rise in energy costs makes further expansion of these disk-based server farms undesirable. Newer, lower-power technologies are desirable.

FIG. 1 highlights a prior-art bottleneck problem with a distributed web-based database server. A large number of users access data in database 16 through servers 12 via web 10. Web 10 can be the Internet, a local Intranet, or other network. As the number of users accessing database 16 increases, additional servers 12 may be added to handle the increased workload. However, database 16 is accessible only through database server 14. The many requests to read or write data in database 16 must funnel through database server 14, creating a bottleneck that can limit performance.

FIG. 2 highlights a coherency problem when a database is replicated to reduce bottlenecks. Replicating database 16 by creating a second database 16' that is accessible through second database server 14' can reduce the bottleneck problem by servicing read queries. However, a new coherency problem is created with any updates to the database. One user may write a data record on database 16, while a second user reads a copy of that same record on second database 16'. Does the second user read the old record or the new record? How does the copy of the record on second database 16' get updated? Complex distributed database software or a sophisticated scalable clustered hardware platform is needed to ensure coherency of replicated data accessible by multiple servers.

Adding second database 16' increases the power consumption, since a second set of disks must be rotated and cooled. Operating the motors to physically spin the hard disks and run fans and air conditioners to cool them requires a substantially large amount of power.

It has been estimated (by J. Koomey of Stanford University) that aggregate electricity use for servers doubled from 2000 to 2005 both in the U.S. and worldwide. Total power for

servers and the required auxiliary infrastructure represented about 1.2% of total US electricity consumption in 2005. As the Internet and its data storage requirements seem to increase exponentially, these power costs will ominously increase.

Flash memory has replaced floppy disks for personal data transport. Many small key-chain flash devices are available that can each store a few GB of data. Flash storage may also be used for data backup and some other specialized applications. Flash memory uses much less power than rotating hard disks, but the different interfacing requirements of flash have limited its use in large server farms. Flash memory's random-access bandwidth and latency are orders of magnitude better than rotating disks, but the slow write time of flash memory relative to its read time complicates the coherency problem of distributed databases.

Balancing workloads among the servers is also problematic. Database server 14 may become busy processing a particularly slow or difficult user query. Incoming user queries could be assigned in a round-robin fashion among database servers 14, 14', but then half of the incoming queries would back up behind the slow query in database server 14.

Improvements in cost, performance, and reliability in data processing systems are made possible by flash memory and other high speed, high density, solid-state storage devices. These improvements are of limited benefit in some scalable cluster systems where data must be partitioned across multiple processing nodes and locally accessed, or placed on a dedicated Storage Area Network, or shared through application inter-process communication. The overhead involved in these existing techniques consumes much of the performance and cost advantage inherent in high density solid-state memory.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the invention are illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

FIG. 1 highlights a prior-art bottleneck problem with a distributed web-based database server;

FIG. 2 highlights a coherency problem with the prior art when a database is replicated to reduce bottlenecks;

FIG. 3 shows a database management system that has multiple instances running in a shared memory space accessing a database that is distributed across flash memory in many nodes according to an embodiment of the invention;

FIG. 4 shows a Sharing Data Fabric (SDF) that enables multiple instances of a DBMS program to access a partitioned database stored in flash memory on multiple nodes according to an embodiment of the invention;

FIG. 5 shows a global, shared flash memory that appears to be a single global address space to multiple servers connected to a sharing data fabric according to an embodiment of the invention;

FIG. 6 shows a hardware node in a global, shared flash memory system according to an embodiment of the invention;

FIG. 7 is a layer diagram of software and hardware layers in a flash memory system using a shared data fabric to enable global sharing of database records in a distributed flash memory according to an embodiment of the invention;

FIG. 8 is a transaction diagram of services and interfaces to a shared data fabric according to an embodiment of the invention;

FIG. 9 shows permanent objects in flash memory being copied to DRAM caches on multiple nodes according to an embodiment of the invention;

FIG. 10 shows components of a DBMS that access a database stored in a distributed shared global flash memory according to an embodiment of the invention;

FIG. 11 is a flowchart of the SDF processing a database record miss in the local buffer cache according to an embodiment of the invention;

FIGS. 12A-D show several alternate interfaces to the SDF according to an embodiment of the invention;

FIG. 13 shows a put record operation that puts a record in the log file for error recovery according to an embodiment of the invention;

FIG. 14 shows an action node requesting a database object from a home node that fetches a modified object on a sharing node using transaction tables and an object directory according to an embodiment of the invention;

FIG. 15 is a snapshot state diagram of a compute node that can act as an action, home, or sharing node for a distributed database program according to an embodiment of the invention;

FIG. 16 shows a get operation inside a transaction that misses in the object cache of the action node, and fetches the object from flash memory of the home node according to an embodiment of the invention;

FIG. 17 shows a commit transaction operation according to an embodiment of the invention;

FIG. 18 shows an abort transaction operation according to an embodiment of the invention;

FIG. 19 illustrates sync, transaction start, and lock operations according to an embodiment of the invention;

FIG. 20 is an illustration of a SDF Agent according to an embodiment of the invention;

FIG. 21 is an illustration of a process structure on the action node and home node according to an embodiment of the invention;

FIG. 22 is an illustration of a process flow of a SDP_GET having a cache hit at the action node according to an embodiment of the invention;

FIG. 23 is an illustration of a process flow of a SDP_GET having a cache miss according to an embodiment of the invention;

FIG. 24 is an illustration of a scoreboard request block format according to an embodiment of the invention;

FIGS. 25-31 illustrate an example of a scoreboard API and scoreboard interaction according to one embodiment of the invention; and

FIGS. 32-36 illustrate synchronization and ordering constructs according to an embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

Approaches for shared global memory accessible to a plurality of processes in a distributed object store that is implemented, at least partially, on flash memory are described. In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the embodiments of the invention described herein. It will be apparent, however, that the embodiments of the invention described herein may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the embodiments of the invention described herein.

FIG. 3 shows a database management system that has multiple instances running in a shared memory space access-

ing a database that is distributed across flash memory in many nodes. Rather than storing the database on a single compute node, the database is partitioned so that different tables are stored on different nodes **242**, **243**, **244**. Reliability is improved since database tables may be replicated and kept updated on several nodes, allowing for backup nodes to replace nodes that fail.

Database tables are stored in flash memory **24** in each node **242**, **243**, **244**, with each node typically storing a different set of database tables. Shared address space **280** (also called a node address space) is used to access flash memory of one or more of the nodes. The nodes provide one or more partitions (also called shards) of the shared address space.

Multiple instances of DataBase Management System (DBMS) program **200**, **200'** are executing within shared address space **280** and are able to access data items in the distributed database that are physically stored in flash memory **24** on nodes **242**, **243**, **244**. Having multiple instances of database management system (DBMS) program **200**, **200'** also improves reliability and reduces bottleneck problems, since user queries may be dispatched to different executing instances of DBMS program **200**, **200'**.

Having DBMS program **200** execute in shared address space **280** allows the program to see just one address space, simplifying DBMS program **200**. Ideally, it would be desirable for multiple executing instances of DBMS program **200**, **200'** running on different physical nodes to observe shared address space **280** so that each could operate as if it is the only executing instance of DBMS program **200**. Thus major modifications and re-writes of the program code of DBMS program **200** could be avoided using shared address space **280**. A DBMS program written for execution on a single address space is preferred since code does not have to be re-written. However, without other facilities, the multiple executing instances of DBMS program **200**, **200'** would contend with each other for the same resources, causing failures.

FIG. 4 shows a Sharing Data Fabric (SDF) that enables multiple instances of a DBMS program to access a partitioned database stored in flash memory on multiple nodes. The SDF is a middleware layer that fits between executing instances of DBMS program **200**, **200'** and nodes **242**, **243**, **244**.

SDF **20** includes an interface for communications between high-level programs such as executing instances of DBMS program **200**, **200'** and lower-level hardware controllers and their software and firmware drivers. SDF **20** is accessible by high-level instances of DBMS program **200**, **200'** using an applications-programming interface (API). Communication between nodes to ensure coherency of database tables stored in flash memory **24** on the multiple nodes is performed by SDF **20**.

Normally, adding nodes provides a less-than-linear performance improvement, since bottlenecks may occur to data stored in just one location on a node, such as shown on FIG. 1. However, using SDF **20**, data records stored in flash memory **24** may be cached near executing instances of DBMS program **200**, **200'** on one or more nodes, allowing multiple processors to access the same data. Coherency of the cached database records is important to prevent data corruption.

FIG. 5 shows a global, shared flash memory that is accessible as a single global address space to multiple servers connected to a sharing data fabric (SDF). Central Processing Units (CPUs) or processors **18**, **18'** can execute programs such as executing instances of a DBMS program to process requests such as user queries of a database that arrive over a network such as the Internet. Each of processors **18** has a cache of DRAM **22** that contain local copies of objects such

as records in a database. These local copies in DRAM **22** are local to the node containing processors **18** and are accessed by processors **18** in response to requests from external users.

While DRAM **22**, **22'** stores transient copies of objects, the objects are more permanently stored in flash memory **24**, **24'**. Objects remain in flash memory **24**, **24'** and are copied to caches in DRAM **22**, **22'** in response to access requests by programs running on processors **18**, **18'**.

Sharing data fabric (SDF) **20** is a middleware layer that includes SDF threads running on processors **18**, **18'**, and APIs and tables of data. A physical interconnect, such as an Ethernet or InfiniBand® fabric, connects physical nodes together. Object copies are transferred across the physical interconnect by SDF **20** from flash memory **24**, **24'** to caches in DRAM **22**, **22'**, and among DRAM **22**, **22'** caches as needed to ensure coherency of object copies.

Flash memory **24**, **24'** can be physically located on many nodes, such as having one flash memory **24** for each processor **18**, or in other arrangements. SDF **20** makes all the objects stored in flash memory **24**, **24'** appear to be stored in a global address space, even though the global address space is shared among many processors **18**, **18'**. Thus flash memory **24**, **24'** together appear to be one global, shared flash memory **26** via SDF **20**. The database is partitioned into many objects, which are stored in a distributed fashion on many nodes within the global shared flash memory.

FIG. 6 shows a hardware node in a global, shared flash memory system. A flash memory system has multiple nodes such as shown in FIG. 6. The multiple nodes are connected together by a high-speed interconnect such as an Ethernet or InfiniBand. One or more links in this high-speed interconnect connect to Network Interface Controller (NIC) **36** on the node shown in FIG. 6.

Processor **18** executes an instance of a DBMS program, threads, and other routines and accesses a local memory that stores program code and data, such as DRAM **22**. DRAM **22** also acts as a cache of objects such as database records in the global, shared flash memory.

Processor **18** also connects to PCIe switch **30**. PCIe switch **30** allows processor **18** to communicate with other nodes through NIC **36** to send and receive object copies and coherency commands. Flash modules **34** contain arrays of flash memory that store permanent objects including database records and tables. Flash modules **34** are accessed by processor **18** through PCIe switch **30**.

FIG. 7 is a layer diagram of software and hardware layers in a flash memory system using a shared data fabric to enable global sharing of database records in a distributed flash memory. Sharing data fabric services **116** include API's that application programs **122** or DBMS database program **119** can use to access objects such as database records and control attributes of the objects. Sharing data fabric services **116** are the API's that communicate with routines and threads in sharing data fabric **112** that provide a unified shared data access of objects including database tables that are permanently stored in flash memory **102**, and may maintain cached copies in DRAM in compute nodes **114**.

Compute nodes **114** are compute nodes, such as node **100** shown in FIG. 6, with processors, DRAM caches of objects, and interconnect. These compute nodes may be constructed from commodity parts, such as commodity processors, interconnect switches and controllers, and DRAM memory modules.

Sharing data fabric services **116** allow application programs **122** and DBMS database program **119** to control policies and attributes of objects by executing routines and launching threads of sharing data fabric **112** that are executed

on compute nodes 114. The exact location of objects and database records within flash memory 102 is transparent to application programs 122 and DBMS database program 119 since sharing data fabric 112 copies objects from flash memory 102 to DRAM caches in compute nodes 114 and may obtain a copy from any location in flash memory 102 that has a valid copy of the object. Objects such as database tables may be replicated to make back-up copies in flash memory 102.

Sharing data fabric 112 performs consistency and coherency operations such as flushing modified objects in a DRAM cache to copy back and update the permanent object in flash memory 102. Sharing data fabric 112 may also migrate flash objects to new flash pages for wear-leveling or other purposes, and update version logs and transaction logs.

Interconnect 110 includes the PCIe switches in each of compute nodes 114, and the high-speed interconnect between nodes, such as Ethernet or InfiniBand links. Sharing data fabric 112 sends objects and coherency commands across interconnect 110 or directly within the compute node, such as directly to flash management 108.

Flash management 108 is activated to migrate flash blocks for wear-leveling and replication. Wear-leveling schemes assign flash blocks for writing in a rotating, least-written, or other fashion to even out usage of flash blocks and prevent early wear-out and failure. Write buffers of flash blocks, logs, and caches may be kept by flash management 108.

Flash interface 106 is an interface between flash management 108 and hardware flash controllers 104, which control low-level access of flash memory 102. While flash memory 102 may have separate modules on different nodes of compute nodes 114, sharing data fabric 112 uses interconnect 110, flash management 108, and flash interface 106 to transparently move objects to and from flash memory 102 on different nodes. Flash memory 102 in aggregate appears to be a single, unified flash memory that is transparently shared among many instances of DBMS database program 119 running on many compute nodes 114.

FIG. 8 is a diagram of services and interfaces to a shared data fabric. DBMS database program 119 could communicate directly with other database nodes using network interface 120, but then the location of objects such as database records in flash memory is not transparent to DBMS database program 119. DBMS database program 119 would then need detailed location information on database records. Send and receive commands to network interface 120 may include commands 128 such as get, put, lock, unlock, start, and abort, which need detailed information on the object's location, such as a file handle or address. Detailed information may include context, thread, container ID, object ID. Location information may be calculated by using a combination of the container ID and the object ID. DBMS database program 119 would have to be re-written to provide this detailed location information, which is undesirable.

Instead, standard, substantially unmodified DBMS database program 119 is used, but instead of using network interface 120 directly, DBMS database program 119 accesses sharing data fabric 112 using API's 116. API's 116 include SDF_GET, SDF_PUT, SDF_LOCK, and other SDF-specific versions of start, abort, commit, savepoint, create, delete, open, and close commands. For example, lock and unlock commands lock and unlock an object using a lock table in sharing data fabric 112 to prevent another user from accessing that object while locked. A node map in sharing data fabric 112 maps objects to address locations in flash memory, allowing sharing data fabric 112 to read and write objects in flash memory through flash management 108 and flash interface 106.

Objects that reside in flash memory on a first node may be accessed over sharing data fabric 112 by sending and receiving messages, and sending object data from a second node over network interface 120. These messages may include commands 128 such as get, put, lock, unlock, start, and abort. These commands 128 are executed by SDF 112 using detailed information on the object's location, such as a file handle or address, that are obtained from a node map, a sharing directory, or a cache map in sharing data fabric 112. Commands 128 and messages are received by a sharing data fabric 112 on the first node, which may access its flash memory to obtain the object. On a read access, sharing data fabric 112 on the first node can then send the object data back through network interface 120 to the second node's sharing data fabric 112.

FIG. 9 shows permanent objects in flash memory being copied to DRAM caches on multiple nodes. Objects may be database records being accessed by a user query. Data fabric switch 40 connects to NIC 36 on three compute nodes. Each node has two NICs 36 to allow for a higher interconnect bandwidth and for redundancy. Each of nodes 50, 52, 54 has a processor 18, flash memory 24, and an object cache in DRAM 22. An Ethernet or other switch (not shown) may also be used, or may be included in data fabric switch 40.

In response to a user database query, a thread executing on processor 18 on node 52 requests access to object 46, which is present in flash memory 24 on node 52. The SDF on node 52 reads object 46 from flash memory 24 and copies the object into its object cache in DRAM 22 as object copy 46'. The DBMS program running on node 52 can then read object copy 46' from its DRAM 22. In this example transfer over data fabric switch 40 was not needed.

In a second example, a thread executing on processor 18 on node 50 requests access to object 47, which is not present in flash memory 24 on node 50, nor in DRAM 22 on node 50. The SDF on node 50 determines that node 54 is the home node for object 47. Node 50 may perform a lookup in a directory to locate the object's home node. The directory may have several parts or levels and may reside partially on local node 50 and partially on other nodes.

An SDF thread on node 50 sends a message to the home node, node 54, requesting a copy of object 47. In response, another SDF thread on home node 54 reads object 47 from flash memory 24 and sends object 47 over data fabric switch 40 to local node 50. The SDF thread on local node 50 copies the object data into its object cache in DRAM 22 as object copy 47'. The DBMS program running on local node 50 can then read object copy 47' from its object cache in DRAM 22.

Object 47 may have already been copied into DRAM 22 on node 54 as object copy 47". Rather than read object 47 from flash memory 24, when object copy 47" is present, object copy 47" may be read from DRAM 22 and then sent over data fabric switch 40 to node 50 to load object copy 47' into DRAM 22 on node 50.

FIG. 10 shows components of a DBMS in more detail that access a database stored in a distributed shared global flash memory. DBMS program 200 is executing on a local compute node that has database records cached in local DRAM as cached objects in local buffer cache 23. There may be multiple executing instances of DBMS program 200 running on different compute nodes, each with its own local cache buffer.

DBMS program 200 has two primary software components that cooperate with each other. Query process 260 receives requests from users that contain a database query. Storage management 270 performs access of database records that are cached in local buffer cache 23. Storage management 270 includes table reader 276 that reads records from portions of a database table cached in local buffer cache

23, table writer 277 that writes or modifies portions of a database table cached in local buffer cache 23, and log writer 278 that logs modifications of records cached in local buffer cache 23. While query process 260 performs query logical operations, query process 260 does not access database records, but instead calls storage management 270 to read or write a record. Separating query and access functions improve software reliability.

SDF 20 is called by storage management 270 when a requested database record is not cached in local buffer cache 23, or during flushes when modified records are copied back to the more persistent storage in flash memory. SDF 20 performs the multi-node operations needed to access data at one of nodes 242, 244 and copy that data into local buffer cache 23 or update that data in flash memory at the node.

The database record may be stored in flash memory at any of nodes 242, 244. Each node 242, 244 has local map 262 that locates the requested data item stored in flash memory at that node. The requested data may be a database record stored in database tables 264, or a database index in database indices 266. Other data stored at a node may include log files 272 or checkpoints 274 that are useful for error recovery.

An example of a Database Management System (DBMS) is a system of one or more software programs, which are written to enable the storage and management of user information in a highly structured and well-defined way. The DBMS enables certain storage properties such as Atomicity, Consistency, Isolation, and Durability, the so-called ACID properties.

Information may be stored as data records organized in n-tuples (also termed simply tuples) of closely related information called rows. A field of a row stores one of the records. Collections of rows are called tables. One or more of a particular field within a table is called a column. A collection of related tables is called a database, and the structure of the related tables is called a schema. Data records may also be referred to as data attributes or data items or objects.

A relational database is a database where information is stored, accessed, and indexed according to specified values of the respective fields, known as a relation. The specified values include specified functions of values, such as ranges of values.

For example, a query is constructed for a relational database which is intended to retrieve only those tuples from the relational database that have a first item record conforming to a first specification in the query AND a second data item conforming to a second specification in the query, where the logical operator AND is also part of the query. Continuing the example, all of the field specifications and the logical operator AND comprise an entity called a relational query specification. In general, a relation refers to the relationship of data items or attributes within a table, or even to the table itself.

The DBMS may be transactional, allowing user operations to be performed as transactions that have well-defined properties. The properties may include an isolation property that ensures that multiple concurrent and distinct transactions operating on the database do not interfere with each other. The transactions each perceive the state of the system as if each of the multiple transactions is the sole transaction executing in the database management system.

Another transaction property is atomicity, meaning that the transaction can be aborted prior to committing any changes to the database management system. The result of aborting the transaction is no change to any record in the database.

The durability property indicates that once a transaction is committed to permanent storage, any database record

changes due to the transaction remain stable even if the system restarts or a power failure or other kind of defined failure occurs.

These properties of transactions may be ensured for a database by a log file in log files 272. A log file is a data structure in which the database management system maintains a sequential record of all data modifications. A log file is used, for example, to record committed transactions and to record abort operations on uncommitted transactions.

When there are multiple users, there may be multiple local buffer caches 23 on multiple nodes. There may be several instances of query process 260 operating on one compute node, using a shared local buffer cache 23, for processing queries by different users.

All local buffer caches 23 and the permanent storage in flash memory of nodes 242, 244 must be kept logically consistent with one another. Periodic checkpoints to flush modified contents of the buffer cache to the permanent storage as checkpoints 274 may be used as one way of maintaining consistency.

FIG. 11 is a flowchart of the SDF processing a database record miss in the local buffer cache. SDF 20 (FIG. 10) is called by storage management 270 when a database record or other data item is not present in local buffer cache 23, step 202. Storage management 270 in the executing instance of DBMS program 200 uses a SDF_GET function in the SDF API to activate a SDF thread that executes on the local compute node's processor, step 204.

The SDF optionally performs a global data lock operation, step 206, in order to ensure that there is a single modifiable copy of the particular data item. The SDF_GET operation may retrieve a modified copy of the particular data item from flash memory or from a local buffer cache on another one of the nodes. A data versioning operation may also be performed to identify the most recent version of the data item, and to save a copy of the most recent version, in case subsequent modifications to the data item need to be rolled back, or un-done. Lock and version operations may not need to be performed in some instances, such as database reads, but may be needed for transactions, such as when writing to database records.

The SDF reads the requested data item from the node, such as from flash memory at a home node for a database record, or from a local buffer cache of a sharing node that has earlier cached the database record, step 208. The requested data item is then loaded into local buffer cache 23 by the SDF, step 210. The SDF may also return a pointer to the data in the local buffer cache so that DBMS program 200 has the pointer.

A cache pin operation may be performed by the SDF, step 212. A cache pin operation ensures that the particular data item remains (is pinned) in the local buffer cache. This guards against another request from this or another thread causing replacement of the data item in cache.

SDF returns a success code to the calling program, such as storage management 270 in DBMS program 200, step 214. Storage management 270 may then access the data item, step 216. Writes to the data item may be allowed.

After DBMS program 200 has finished accessing the data item, step 216, DBMS program 200 calls the SDF to unlock the data item, step 218, if that data item was previously locked in step 206. When the data item was pinned by a cache pin operation in step 212, then DBMS program 200 calls the SDF to perform a cache unpin operation, step 219.

When the data item was written in step 216, the DBMS program optionally calls the SDF to flush the modified data item back to the flash memory at its home node, step 224. The SDF then stores the modified data item at its home node, step 226. Various logs may need to be written by the DBMS

11

program or by the SDF to ensure that this write back is durable. The DBMS program optionally calls the SDF to perform a SYNC operation, step 220, so that the modified data item is made visible to other nodes, step 228. The SDF_SYNC operation makes the results of selected previous operations including the modified data item visible to other instances of DBMS program 200' that are executing on the local node or on other compute node in the system. A global cache directory at the home node may be updated to make the modified data item visible to other nodes.

Other database operations, such as row insertions, row deletions, and index updates, are performed in a similar manner using the SDF API.

FIGS. 12A-D show several alternate interfaces to the SDF. In FIG. 12A, SDF 20 has an API that uses multi-processing interfaces such as SDF synchronization operations in place of synchronization primitives typically provided by a multi-processing system. The SDF API has software libraries with interfaces that are compatible with existing multi-processing interfaces, such as Symmetric-Multi-Processing SMP interfaces. Since the SDF API has functions with interfaces matching the SMP interface, programs written for the SMP interfaces require little or no modification to operate with SDF 20. The SDF synchronization operations are enabled to operate transparently across multiple nodes in a cluster. Data items stored in flash memory 24 or in buffer caches 222 of nodes 242, 243, 244 can be accessed by DBMS program 200 using these SMP-compatible interfaces that cause SDF 20 to copy the data items into local buffer cache 23, and flush modified data items back.

FIG. 12B shows a SDF interface that uses a file system interface. DBMS program 200 uses data containers rather than files or magnetic disk partitions. Database tables, records, or other data items are stored in data containers 252, 253, 254 on nodes 242, 243, 244, but appear to be local to DBMS program 200. DBMS program 200 is written to advantageously use functions of SDF 20 that mimic file-system functions for magnetic disk partitions, such as a lseek() function.

DBMS program 200 accesses data items stores in data containers 252, 253, 254 using SDF 20's API, which is compatible with existing DBMS access methods using file systems and/or device partitions. Durability and consistency of data stored in data containers may be automatically and transparently maintained by SDF 20.

FIG. 12C shows a SDF interface using SDF MUTEX operations. Software compatibility libraries 248 that DBMS program 200 is written for include a symmetric multi-processing (SMP) mutual-exclusion (MUTEX) operation. When DBMS program 200 calls the SMP MUTEX operation in compatibility libraries 248, a SDF_MUTEX operation is called in its place. The SDF_MUTEX operation has the same list of parameters and variables in its interface, and thus is transparent to DBMS program 200.

SDF 20 allows direct substitution of data fabric primitives such as SDF_MUTEX for similar primitives such as SMP_MUTEX provided by platforms supporting a single address space. For example, SDF 20 provides a test-and-set operation which is used to create a MUTEX operation that is a direct replacement for an SMP-based MUTEX operation.

Programs written for the SMP interfaces require little or no modification to operate with SDF 20. The MUTEX synchronization operations are the enabled to operate transparently across multiple nodes in a cluster.

FIG. 12D shows a SDF interface using a container of synchronization primitives. DBMS program 200 is written to use synchronization primitives such as test-and-set, and com-

12

pare-and-swap, which consist of multiple operations performed atomically on one or more program variables. These variables are enabled to be shared among multiple instances of DBMS program 200 by placing them into primitive operation container 240, which is shared among the multiple instances. SDF 20 allows DBMS program 200' (not shown) executing on other nodes to access primitive operation container 240.

Special synchronization operations such as test-and-set, and compare-and-swap, are implemented efficiently within SDF 20 and provided to DBMS program 200 through the SDF API, which has special compatible functions for these operations.

Since the SDF API has functions with interfaces matching the interface of these primitive operations, programs written for the primitive-operation interfaces require little or no modification to operate with SDF 20. The existing calls to the synchronization operations are redirected to use objects in the new synchronization container. The SDF synchronization operations may be enabled to operate transparently across multiple nodes in a cluster.

FIG. 13 shows a put record operation that puts a record in the log file for error recovery. SDF 20 may support distributed log file operations through log file containers. A log file container may use a special node address space for performing functions of a distributed log file. A log file container may maintain certain attributes important to log files, such as n-way replication. The log file container may allow multiple instances of DBMS program 200, 200' to commit database updates in a global order.

For example, a DBMS storage manager instance inserts a record into log file container 250. In response to a request from a transaction executing in any DBMS program instance, the DBMS storage manager instance performs a data fabric PUT RECORD operation to add a log record to log file container 250. The SDF performs the operation, updating the contents of log file container 250 and updating internal pointers. This operation may be done atomically by the SDF and be recoverable by the SDF when failures occur in the hardware or software.

The SDF may perform replication operations to replicate the updated record, and wait for the log record to be permanently stored before returning a success code to the DBMS storage manager instance. Each PUT RECORD operation may be associated with a Logical Sequence Number (LSN) from LSN generator 246. DBMS program 200, 200' is able to subsequently use the LSN to retrieve the respective log record should a database recovery or rollback operation be necessary. Various techniques for logging and recovery may use the log file container 250.

FIG. 14 shows an action node requesting a database object from a home node that fetches a modified object on a sharing node using transaction tables and an object directory. A node may operate as one or more of a home node, an action node, or a sharing node for any particular object. Objects include database records or other data items associated with a database.

Home node 52 is the permanent location of object 46 in flash memory 24. Action node 50 is executing an application program that requests an object. Sharing node 54 has a copy of the object in its object cache in DRAM 22.

A program executing on action node 50, such as executing instances of DBMS program 200 (FIG. 10), requests access of object 46. Object 46 is not yet present in DRAM 22 of action node 50, so the SDF determines the object's home node, such as by hashing the object's name or identifier or looking up the object in a table.

Transaction table 42 is stored in DRAM 22 of action node 50 and eventually contains information to identify home node 52. In response to a request from action node 50 to access object 46, the SDF on home node 52 looks up the object in its object directory 44. Object directory 44 indicates that although object 46 is present in flash memory 24 of home node 52, this object 46 is stale. A modified object 46' is present in DRAM 22 of sharing node 54 in this example.

Since modified object 46' is more current than stale object 46 in flash memory 24, SDF on home node 52 sends a message to SDF on sharing node 54. This message causes transaction table 42 on sharing node 54 to be checked to make sure that modified object 46' is not locked for use by a program executing on sharing node 54. If modified object 46' is locked, action node 50 waits until the lock is released. Release of the lock causes an update of transaction table 42 on sharing node 54.

When transaction table 42 indicates that modified object 46' is unlocked, SDF on sharing node 54 sends the object data of modified object 46' over data fabric switch 40 to action node 50. Object copy 46" is created in DRAM 22 on action node 50. The requesting program on action node 50 can now access object copy 46".

Other steps may be performed to update object directory 44 and stale object 46 in flash memory 24 on home node 52, although this may be delayed to allow home node 50 to update object copy 46" to reduce the number of writes and operations to flash memory 24.

FIG. 15 is a snapshot state diagram of a compute node that can act as an action, home, or sharing node for a distributed database program. Node 100 is one of compute nodes 114 of FIG. 7, with hardware such as detailed in FIG. 6.

Node 100 has threads running on processor 18, including application thread 90 and/or SDF thread 92. Application thread 90 can be an executing instance of DBMS program 200 on local node 100.

Messaging 94 allows SDF thread 92 to send and receive messages from other nodes. Messaging 94 may use software and hardware such as interconnect 110 of FIG. 7, NIC's 36 of FIGS. 6, 9, and other hardware such as switches.

Node 100 includes DRAM 22 and flash memory 24. The DRAM state shown is a snapshot of the contents of DRAM 22 at a particular point in time, while the flash state is a snapshot of the contents of flash memory 24 at approximately that same time.

Flash memory 24 at each node stores flash objects 80, which may be grouped into containers. A flash object may be uniquely addressable in the SDF using a container-identification and an object identifier. Metadata 88 stored in flash memory 24 may include container metadata and global metadata. Container metadata describes the attributes or properties of objects in a container (such as a number of replicas for the container). Global metadata may include virtual-to-physical node-identification mappings and/or hash functions. The hash function is performed on the object's name to generate an identifier that can be used in further table lookups to locate the object using a global map.

Object cache 84 in DRAM 22 stores copies of objects that are also stored in flash memory 24 of the home node, which may be node 100 or may be another node. DRAM objects 86 are objects that have node 100 as their home node, but the object's attributes specify that the object primarily resides in DRAM 22 rather than in flash memory 24. For example, frequently-accessed objects such as database indices 266 of FIG. 10 may be selectively enabled to remain in DRAM 22 rather than the flash memory to improve look-up performance.

DRAM 22 also stores state information for particular classes of nodes. Action node state 70 stores state information for objects using node 100 as the action node, while home node state 72 stores state information for objects using node 100 as their home node, such as DRAM objects 86 and flash objects 80.

A home node is a well known, an authoritative source of the object, which resides in DRAM, flash memory, or another component in the storage sub-system. While an object may be cached and/or replicated in DRAM and/or flash memory at several nodes, only one of these nodes (at any one time) is considered to be the home node for that object. An action node stores a transient copy of an object. The action node usually obtains a copy of the object from the home node. Node 100 can operate as the home node for some objects, and as the action node for other objects.

Transaction table 74 in action node state 70 stored in DRAM 22 has entries to track transactions. Transaction table 74 keeps a list of all objects accessed (touched) during a transaction, and may keep a copy of the object's initial state when the object is modified, or a pointer to a log of changes from an initial state that is archived. The initial state pointed to by transaction table 74 aids in object recovery if an abort occurs. Transaction table 74 provides recoverability for threads of applications and SDF services that execute on the node's processor. There may be links between tables, such as links in entries in transaction table 74 to entries in transient protocol state table 76.

A transaction is a series of operations. The transaction as a whole succeeds as a unit or the transaction and all operations in the transaction are aborted. This may also be referred to as an atomic set of operations. In a transactional system that maintains isolation among concurrent transactions, there are no unexpected effects from an aborted transaction since either all operations in the transaction are executed completely, or any partially-executed operations of the transaction are aborted without visible side effects. Transactions are extremely useful for distributed database programs, since a database record may be updated as a transaction, preventing a partial update of that database record.

Transient protocol state table 76 in action node state 70 stored in DRAM 22 has entries to track outstanding requests. The requesting thread, type of request, request status, and order of requests from each thread are stored in this table, which is shared by all threads at node 100.

Home node state 72 stores information for objects that have node 100 as their home node, and are thus stored (homed) in DRAM 22 (as DRAM objects 86) or flash memory 24 (as flash objects 80) of node 100. Home node state 72 in DRAM 22 stores a sharing list within global cache directory 78, and flash object map 82. Global cache directory 78 stores the state of all objects that have node 100 as their home node and have been cached in one or more nodes in the system. The state of the object in global cache directory 78 is one of shared, modified, invalid, and either locked or unlocked. An indication of whether the object is being accessed within a transaction may also be stored in global cache directory 78. The locations of cached copies at other nodes, and the state and progress of any coherency operations for the objects are also kept in global cache directory 78.

Flash object map 82 maintains a map between an object identifier and its location in flash memory 24. Flash object map 82 may store an address, a partial address, or an identification of the location of the object in flash memory 24. Flash object map 82 maps at least a portion of a respective identifier of a particular one of the objects to a physical location in flash memory 24. Flash object map 82 has entries for all flash

15

objects **80** that belong to this home node, regardless of whether that object is cached at any node. Flash object map **82** may be a homed object map that also has entries for all DRAM objects **86** at this home node.

Database Transaction Flows Using SDF—FIGS. **16-19**.

Snapshots of the states and movements of database objects and SDF messages among two nodes in a multi-node shared flash memory system are shown in the examples of FIGS. **16-19**. Snapshot diagrams, similar to that of FIG. **15**, are shown for action node **50** and home node **52**.

Extensions of these flows could involve sharing node **54** (FIG. **14**) when sharing node **54** contains a modified copy of the requested object. Then the requested object is sent from sharing node **54** rather than from home node **52**, since the flash object at home node **52** is stale. Messages passed among these nodes, including the database object being copied, are shown by the arrows.

These examples are for operations that are part of transactions. To provide ACID properties to executing instances of DBMS program **200**, **200'**, operations that access database records are organized as transactions. Transactions allow the operations of a transaction to be committed together as an atomic unit, preventing partial updates of database records that can corrupt the database. Operations that are not part of transactions could also be used for less important accesses, such as status inquiries or database reads. When a requested object is already present in object cache **84** of action node **50** (a hit), application thread **90** may simply read the object from object cache **84**.

FIG. **16** shows a get operation of a transaction that misses in the object cache of the action node, and fetches a requested object from flash memory of the home node. A miss occurs when application thread **90** checks its local DRAM object cache **84**. The node that application thread **90** is running on is referred to as action node **50** since actions are being performed at this node. A request is sent from the action node to the home node for the requested object. The location of home node **52** is determined by action node **50**, such as by hashing the name of the object being requested and using the hash as an address in lookup tables. The lookup tables may be a global map from identifiers of the objects to home nodes of the objects.

Application thread **90** uses the address or identifier for home node **52** to send a message to home node **52**. This message requests the object from home node **52**. At home node **52**, the message received from action node **50** activates SDF thread **92**, which looks up the object identifier in global cache directory **78** at home node **52**. In this example, no copies of the object have been cached by any other nodes, so a directory miss occurs.

SDF thread **92** running on home node **52** then looks up the object identifier in flash object map **82** to find the address of the object in flash memory **24** of home node **52**. Flash memory **24** is read to copy flash object **80** stored in the flash memory of home node **52**. A copy of flash object **80** is sent from SDF thread **92** at home node **52** to application thread **90** at action node **50** via an interconnect between home node **52** and action node **50**. Application thread **90** (or a SDF miss-handling thread such as SDF thread **92** on action node **50** invoked by application thread **90**) then loads the copy of the object into object cache **84** at action node **50**.

Transaction table **74** is updated to include an entry for the copy of the object that was just loaded into object cache **84**. This entry identifies the current transaction that requested the object. At the start of a new transaction, application thread **90** can create a new entry in transaction table **74**, and this entry is updated with a list of objects touched by the transaction as the

16

transaction is processed, or with pointers to sub-entries for each object in that transaction. Application thread **90** can resume processing the transaction and read the object copy in its object cache **84**.

FIG. **17** shows a commit transaction operation. Once all of the operations in a transaction have been started and are nearing completion, the transaction ends using a commit transaction operation as shown in FIG. **17**. The commit transaction operation ensures that all operations of that transaction have been successfully completed before any modified objects such as updated database records are committed to persistent memory. Thus the transaction is committed as a whole, all-or-nothing. All objects modified by the transaction are updated, or none are updated.

When application thread **90** reaches a commit transaction operation, application thread **90** reads transient protocol state table **76**. Transient protocol state table **76** contains a list of all outstanding requests for all prior threads at action node **50**, the status of all requests, and the order of requests for each thread. Application thread **90** waits until completion of all outstanding requests for the current transaction for application thread **90**. If there are any dependencies among threads, application thread **90** must wait for completion of dependent requests at other threads, according to any ordering rules.

Once all dependent outstanding requests have completed, as indicated by transient protocol state table **76**, application thread **90** reads the transaction's entry in transaction table **74**. A list of all objects touched by that transaction is read from transaction table **74**. Objects that were only read do not need to be copied back to the home node, but modified (dirty) objects do need to be copied back.

Each of the modified objects for this transaction are sent back to their respective home nodes, or flushed. A flush operation causes a cached object to be sent to home node **52** in order to synchronize the most up-to-date state of the object with the source. A flush to the object source in flash-memory provides persistence and a level of durability to the object state.

A flush may not require that flash memory is immediately written with the modified object. Instead, the modified object may be stored in DRAM or a write buffer on home node **52** when the flush is completed. Later, home node **52** may perform the actual writing of the modified object to flash memory.

A two-phase commit may be used to avoid contentions with other nodes that may also be accessing one of these objects at about the same time. Action node **50** may in a first phase indicate a desire to write a modified object back to the home node, and receive a timestamp, and then in a second phase actually write the modified object if there are no objections from other nodes. If another node objects, such as by also indicating a desire to access the same object, the timestamps can be compared, and the node with the earlier timestamp wins.

After the two-phase commit process has succeeded, SDF thread **92** on home node **52** locates the homed object inside flash memory using flash object map **82**, and the modified object from action node **50** is written into flash memory as one of flash objects **80**. Global cache directory **78** may first be consulted to verify that no other nodes have this object, and invalidations may be sent to any sharing nodes.

Global cache directory **78** is updated to indicate that action node **50** no longer has this object locked. SDF thread **92** on home node **52** sends a message to action node **50** to unlock the modified object that was just updated at home node **52**, and application thread **90** on action node **50** unlocks the object in object cache **84**. The object could be deleted from object cache **84**, or changed from modified to shared, and changed to

17

the unlocked state to indicate that this object cannot be written until a new lock is obtained from home node 52.

The transaction's entry in transaction table 74 is deleted once all modified objects have been successfully flushed to their home nodes, and unlocked in object cache 84. The transaction is finished and has been committed.

If any modified object cannot be written back to its home node, such as if the home node crashed, then the transaction being committed must be aborted. Any modified objects that have already been written back to their home nodes must be restored to their initial conditions. Log files may be used to recover from this rare situation.

FIG. 18 shows an abort transaction operation. Sometimes, one of the operations inside a transaction cannot be completed, or has an unsuccessful return code, perhaps after several retries. A requested object may be locked by another node and unavailable, and that other node may have crashed or locked up, preventing release of the requested object. An abort transaction operation may then be performed.

Application thread 90 reads transient protocol state table 76 and waits for all outstanding dependent requests that are ordered before this transaction to complete. Then transaction table 74 is read to obtain a list of all objects touched by the transaction being aborted. Transaction table 74 contains the initial states of all objects in the transaction, or pointers to these states, or other information that allows the initial states to be obtained or generated. For example, the initial state of an object may be stored on the home node of that object.

All touched objects in this transaction are restored to their initial state at the beginning of this transaction, such as by restoring objects in object cache 84 using the initial states from transaction table 74. Alternately, each of the touched objects may simply be invalidated in object cache 84.

Restored objects in object cache 84 that were locked by the aborting transaction are unlocked, with an unlock message being sent to the home node for each object being unlocked. Home node 52 updates global cache directory 78 to indicate that the object is unlocked. Other nodes may now access the object. Once all restored objects have been successfully unlocked, the entry for the aborting transaction can be removed from transaction table 74.

FIG. 19 shows sync, transaction start, and lock operations. A transaction start operation creates a new entry in transaction table 74 at action node 50. A copy of all objects that will be touched by the new transaction are stored in transaction table 74 or otherwise archived to allow rollback to this initial state should the new transaction later be aborted.

A sync operation, such as an SDF_SYNC, is performed to synchronize application thread 90 with other threads on action node 50 or on other nodes. Application thread 90 (or SDF thread 92 if called by application thread 90 to perform the sync) reads transient protocol state table 76, which has an ordered list of all outstanding requests for all threads at action node 50, and the status of those requests. Application thread 90 waits until all outstanding dependent requests have completed.

The sync operation uses a set of ordering rules to determine which outstanding requests must be completed prior to the sync operation completing. In some embodiments, each sync operation is enabled to select the ordering rules it uses. In a first example, a sync operation executed by an application does not complete until all outstanding operations of the application have completed. In a second example, a sync operation does not complete until all outstanding write operations of a particular transaction have completed. In a third

18

example, a sync operation does not complete until all outstanding operations of the action node performing the sync operation have completed.

Once all outstanding requests, as determined by the ordering rules in use, have completed, the sync operation is completed. Application thread 90 can resume having synched to other threads.

A lock is requested before application thread 90 writes to an object that has already been loaded into object cache 84 on action node 50. Once locked, other nodes cannot write that object.

Application thread 90 sends a lock request message to home node 52. SDF thread 92 on home node 52 looks up the object's entry in global cache directory 78 and waits until the object is available and not locked by any other thread on any node. Once the object is free, SDF thread 92 on home node 52 returns the lock to action node 50.

An ordered queue may be used to process lock requests at home node 52. In addition to basic mutex-like single state locks, read-write, and upgrade locks, various complex sets of locks may also be implemented (e.g. multi-granularity and hierarchical locks).

Several other embodiments are contemplated by the inventors. For example, while PCIe switch 30 has been described, other local buses could be used, and switch 30 could be a HyperTransport switch rather than a PCIe switch. Multi-protocol switches or multiple switches or hubs could be used, such as for supporting HyperTransport and PCIe on the same node. Data fabric switch 40, PCIe switch 30, and interconnect 110 may have overlapping hardware or software and operate to allow messages to pass for SDF.

Rather than using a local-bus switch, other network topographies could be used, including rings, chains, hubs, and links. Although flash memory has been described, other solid-state memories could be used for storing the objects at the home node (home objects), such as phase-change memory, ferroelectric random-access memory (FRAM), Magnetoresistive RAM (MRAM), Memristor, Phase-Change Memory (PCM), Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) memory, Resistive RAM (RRAM), Racetrack memory, nano RAM (NRAM), and other non-mechanical non-volatile memories. Flash memory uses electrically-erasable programmable read-only memory (EEPROM), although other technologies such as Phase-change-memory (PCM) may be substituted. NAND flash memory may be used to provide advantages in performance, such as write bandwidth, compared to other non-volatile, electronic memory types. Additional levels of storage hierarchy could be added, such as hourly, daily, or weekly backups to disk, tape, or optical media. There could be many flash modules or only one flash module.

While the description herein may describe the global, shared flash memory as being accessible in one global shared address space, in other embodiments, the global, shared flash memory is accessible in a plurality of global address spaces. For example, in some embodiments, each container is accessible by a respective address space.

The Sharing Data Fabric (SDF) is a unified user-space mechanism to access and store data into hierarchical DRAM, flash memory and the storage sub-system of a clustered or distributed set of compute nodes. SDF uses user-defined attributes to control access, consistency, duplication, and durability of objects in storage. To each application executing on any compute node, the distributed data and storage appears to be logically one big device with integrated cache, memory and storage.

The layers of software and hardware in FIG. 5 may use various combinations of hardware, firmware, middleware, drivers, software, etc. and the layers may be modified in various ways. The connections and interfaces between layers and components may also vary from the simplified diagrams of FIGS. 5, 6. Executing instances of DBMS program 200, 200' may operate on multiple nodes, with one instance per node, or many instances per node. Several query processors could share a common storage management 270, or each query processor could have its own storage management 270. Many other arrangements and partitionings of blocks are possible. DBMS 200 may be substantially unmodified, yet be relinked or use a different library, or may change some routine names, and may changing how locking is done.

When transactions are not supported or used, transaction table 74 and (optionally) transient protocol state table 76 may be omitted. Other tables, lists, or data structures may be used to track SDF operations at the action and home nodes. Tables may contain entries in table format, or as linked lists, or in other formats, and can be flat, hierarchal, multi-level, or in a variety of other formats. Global cache directory 78 may contain sharing lists with or without other information.

Transient protocol state table 76 in action node state 70 stored in DRAM 22 has entries to track outstanding requests. Rather than storing information on threads, information on contexts may be stored in state table 76. The requesting context, type of request, request status, and order of requests from each context are stored in this table, which is shared by all contexts and their threads at node 100. An indirection of "context" is used to link a sequence of activity of gets, puts, etc. An application thread can use multiple contexts, or multiple threads can use one context. Application threads cannot see any SDF related tables, only SDF protocol threads can. By using contexts, monitoring of what application thread is calling which SDF calls is not needed. This makes the API more flexible.

For FIGS. 16-19, an asynchronous messaging model could be enabled by activating a receiving SDF thread 92 at action node 50 when a message returning an object copy is received from home node 52. Then the return arrow would go to SDF thread 92 rather than to application thread 90 at action node 50.

In this variation, receiving SDF thread 92 then loads the object copy into object cache 84 of action node 50 and application thread 90 can use the object copy. This handoff using the receiving SDF thread isolates application thread 90 from the details of MPI messaging and may improve robust multi-threaded execution.

While a database program requesting an object has been described, other kinds of programs such as networked services, applets, proxies, clients, servers, etc. may request objects and operate in a manner similar to that described for application programs 122. Each node could run one application program such as a server application, or multiple programs of the same or differing types. These programs may themselves perform some caching of data. Some applications or networked services may bypass SDF and reach the network interface directly, or may do so for some kinds of accesses that do not require SDF. Other kinds of API calls and network procedures or calls may be used than those listed in FIG. 8, and additional API functions may be added. Different kinds of messaging between nodes may be employed other than MPI or MPI-like messaging.

While computing nodes have been described as each having a processor, DRAM cache, flash memory, and a NIC, some nodes could be compute-only nodes without any flash memory. Other nodes may be storage-only and have flash

memory but do not execute application programs 122. Nodes may have more than one processor, and may have additional memories, such as a read-only memory for program code, static random-access memory (SRAM), or other DRAM. Several levels of processor cache may be present that may keep even more transient copies of objects in DRAM 22. The processor may include one or more processor chips, which each may have one or more processor cores. For example, in some embodiments the processor includes two, quad-core AMD Opteron™ processor chips.

A computing node may have a processor that executes both a web server and a database server application, or a combined application or applet. The compute node may be able to execute several such applications simultaneously, or several threads that can each handle one or more remote user requests. Software may be partitioned or combined in many different ways. In a first example, some or all of the SDF API's are statically linked into an application program. In a second example, some or all of the SDF API's are in a library that is dynamically linked into an application program. In a third example, some or all of the SDF API's are integrated with an operating system. In a fourth example, a stub library is used to replace at least some operating system calls (such as for file operations) with calls to API's in an SDF library.

The NIC may connect to an Ethernet, an InfiniBand, or some other kind of network, or even multiple network types. While two NIC's 36 are shown per node in FIG. 14, one NIC 36 may be used per node, or more than two. Nodes may be asymmetric and have different hardware rather than all be identical. In FIGS. 16-17, the homed object may be stale, and a more recent modified object from a third-party sharing node may be fetched instead. An acknowledgement to home node 52 then may come from the sharing node rather than from action node 50. Other flow modifications are possible.

In some systems, compute nodes may have multiple processors that have separate DRAM caches but share flash memory. Groups of logical compute nodes may use the same physical hardware. One compute node could act as both action node 50 and home node 52 for one object in some of these variations. SDF could send messages between processors on the same compute node that are acting as action node 50 and home node 52. These intra-node messages may or may not use NIC 36.

Direct-memory access (DMA) may be used to transfer blocks of data, rather than using I/O or other instructions by processor 18 (FIG. 6). The terms permanent and transient are relative to each other rather than absolute. Transient objects in DRAM caches may be flushed or copied back to flash periodically, or after a period of no accesses. DRAM caches may be organized per node, per application, per thread, per container, and various other organizations. A permanent object in flash may still be moved from one node to another, or from one flash block location to another at the same node, such as for wear-leveling, data migration, or de-fragmenting purposes. Permanent is meant to imply durability, or retention of data upon power loss, rather than an unlimited life-time of the object.

A particular object may be replicated, so that several copies of the object are located in flash memory of several different nodes. This replication improves reliability since the home node can fail and one of the back-up replica copies may then be used as the home object. At any given time, only one of the replicas may be designated as the permanent object in flash, while the others are kept as backups. The replica copies may provide additional read-only access to the object.

The node chosen as the home node could change over time to spread the wear of the replicas. Multiple replicas could be

used as home objects, but for different groups of nodes, so that each group of nodes has only one of the replicas as its home for the object. In some embodiments, replicas provide read-only access, and write access is performed at the home node.

While the permanent object has been described as residing in flash memory of the home node, and transient copies of the object have been described as residing in DRAM cache on any node, some classes of object may have varying properties. For example, some objects may be stored primarily in DRAM rather than in flash memory, such as DRAM objects **86** of FIG. **15**. Some objects may be DRAM-only objects that are never stored in flash memory. Instead, the permanent object is in DRAM at the home node. Alternately, some objects may not be allowed to be present in DRAM caches, but only in flash memory at the home node. These flash-only objects may be copied to DRAM for a specified short period of time only. Some objects may have time limits for residing in DRAM, or may specify how often they must be copied back to flash memory. The maximum number of shared copies allowed may be specified as an attribute.

Objects are a type of element that can be stored in a container. Elements can be fixed-sized or variable-sized. Variable-sized elements may be termed objects. The description above applies similarly to objects that are fixed-size elements, with some differences such as an ability to access fixed-size elements using, for example, a table look-up (based on an address of the element) rather than a hash (based on a key of the object's name).

To aid consistent distribution and location of data through the clustered or distributed flash-memory system, metadata **88** (FIG. **15**) may include an indication of a hash function to be performed on the object identifier to generate an identifier of a unit of storage known as a shard, and an indication of a hash function on the container name to generate a node name. A shard-mapping table maps shard identifiers to nodes (for example, via virtual node names), and another table may be used to map virtual node names to physical node addresses to provide a level of indirection in order to enable dynamic configuration of the distributed flash-memory system. Flash object map **82** may be a node map or a local map. Metadata **88** may be stored in flash memory **24** or may be stored in another memory such as a non-volatile write buffer or a battery-backed memory.

In addition to the threading model described where each request is handled by a thread, an event based model could also be used where contexts (with state and progress indicators) are maintained for each request utilizing a handful of threads. Application thread **90** in FIG. **16** uses the address or identifier for home node **52** to send a message to home node **52**. Alternatively, at any time during processing of an outgoing request, the application thread may hand off the request to the SDF thread, which performs any of the determining or messaging functions on behalf of the application thread.

The word "may" indicates optionally and/or selectively. An object name may be an object identifier, and an identifier can be a name, key or address. The term thread is sometimes used generically to refer to a locus of execution, and may refer to one or more threads, processes, programs, applications, applets, objects, executing contexts, etc.

In addition to storing all tables in DRAM, tables could also be stored in a DRAM and flash-memory based hierarchy. Tables could be backed up to flash periodically. Backing up tables to flash memory could be specific to a local node or global. In addition to application thread **90** executing SDF protocols on the action node, SDF may process the protocol on a thread that may process more efficiently and act as a proxy for application thread **90**. Such a thread may reside in

the application process or in an SDF daemon on the same node. Various arrangements, partitionings, and levels of threads, programs, routines, and applications are possible.

The techniques described herein are applicable to various types of databases, such as row-oriented databases, column-oriented databases, relational databases, transactional databases, and databases with and/or without some and/or all of the ACID properties. While specific examples of SDF commands have been given, there are, of course, many possible ways of arranging and/or constructing and/or providing one or more data fabric commands and/or other ways of controlling and/or managing a data fabric to achieve similar effect. Transactions may include lightweight transactions such as for locks, and persistent transactions.

The specific API functions provided by the data fabric vary according to different embodiments. Standard APIs used by applications executing in a single address space are replaced with data fabric APIs that transparently enable the application to execute in a distributed fashion as shown in FIGS. **12A-D**.

Flash interface **106** and flash management **108** (FIGS. **7-8**) may be software, firmware, hardware, or various combinations. Hardware flash controllers **104** may include firmware. Primitive operation container **240** (FIG. **12D**) may also contain log files, database indices, and database tables.

A global ordering of operations on the database may be implemented with an ordering protocol. A protocol for coordinating operations from multiple nodes to the same database in order to provide a global ordering, is two-phase commit. In two-phase commit, there are two communication phases between a particular set of nodes for every database update, the first phase to "prepare" the set of nodes and agree on the update, the second phase to perform the update. A logged 2-phase commit process may be used to facilitate failure recovery. Recoverable decision and update schemes for a clustered or a distributed system such as 3-phase commit, voting or consensus may be used in place of 2-phase commits.

Flash memory based log file containers are enabled to store log records in non-volatile memory, such as battery-backed DRAM, accessible via a flash memory controller, and return the success code with DRAM latency vs. flash memory write latency. For example, the latency to store a 256-byte record to non-volatile DRAM memory is less than 10 microseconds. By comparison, storing 256 bytes to some flash memory takes at least 200 microseconds, and could take longer, if small writes such as 256-byte writes are not directly supported by the flash memory. Other kinds of files, records, or containers could also be stored in a similar manner.

The SDF data fabric further enables efficient checkpoint operations. One or more buffer caches, such as one buffer cache per node, are managed by the data fabric for application programs to use. As data items, such as attributes, database records and rows, are modified, the data items are optionally and/or selectively kept in one of the buffer caches rather than being flushed back to permanent storage. The data fabric may be enabled to provide a copy of a modified data item in the buffer cache of one of the nodes to a program, such as a DBMS storage manager instance executing on another of the nodes and performing a data fabric access operation, such as GET, in lieu of fetching a (stale) copy of the data item from permanent storage.

Periodically, a DBMS program may flush some or all of its modified items back to permanent storage, such as flash memory, to keep the permanent storage contents relatively current. The data fabric may be enabled to accelerate the flushing by automatically copying all of the modified data

items back to permanent storage. For example, the DBMS software may use a data fabric GLOBAL FLUSH command to initiate this operation.

DBMS program 200, 200' are scalable, since additional compute nodes may be added, each with a DBMS program 200, to improve the number of user queries that can be processed simultaneously. All compute nodes are able to access the same partitioned database that is stored persistently in flash memory on the storage nodes. The SDF allows all compute nodes to access the same database on the storage nodes in a scalable fashion, since SDF caches portions of the database being accessed in each compute node's local buffer cache. The performance and query throughput is a linear (or nearly linear) function of system cost and/or size (number of nodes). The system can be more cost effective and/or scale to larger sizes than symmetric multiprocessor (SMP) systems. A DBMS which has been written to execute in a single address space can execute as DBMS program 200 on the scalable hardware and software platform shown in FIGS. 5-8 with minimal or no modifications of software of the DBMS. Certain centralized functions of a DBMS, such as the logging and the checkpointing, are efficiently scalable (and thus efficient for larger systems) without extensive modifications to the DBMS software.

In an embodiment of the invention, a system comprises a number of processing nodes, each node optionally having one or the other or both of the ability to run application programs and the ability to manage some amount of flash memory and/or other type of high density, solid-state memory. These various types of high density, solid-state memory are referred to herein as flash memory without being interpreted as limiting the scope or applicability of the techniques presented herein.

Advantageously, embodiments of the invention provide for a system in which application processes are executable on any processing node in a cluster, and the application processes are further executable on more than one node, such as for load balancing purposes. Embodiments of the invention further provide for flash memory that is be globally accessible by an application running on one or more nodes in the cluster.

Advantageously, embodiments of the invention also provide for flash memory that is accessible by applications without the overhead of Input/Output subsystem operations, as is the case typically with, for example, magnetic disk storage. This overhead consumes time and CPU cycles and wastes storage and interconnect bandwidth due to the large fixed-block-sized operations. A lightweight access method is desired to allow the performance potential of flash memory to manifest in scalable cluster systems.

Embodiments of the invention also provide for mechanisms for data access that are either synchronous (meaning that the application program waits for completion), or asynchronous (meaning that the application proceeds in parallel with the operation being performed in the system, and determines at some later point if the operation has been performed, waiting for its completion if desired).

Advantageously, embodiments of the invention also provide for underlying shared memory mechanisms, which implement the lightweight access mechanism, and other system tables and data structures that are robust in the presence of failures, for example a power loss affecting one or more nodes, or a software crash on one or more nodes. In certain embodiments, when some nodes fail, other nodes continue to use the remaining system resources. Methods for resource partitioning and replication are enabled, by use of techniques presented herein, to implement a high availability and/or fault tolerant system.

Advantageously, embodiments of the invention also provide the communication mechanisms which are configurable such that one or more of nodes, flash memory, and application processes can be added and/or removed from the system without interrupting the operation of the nodes, flash memory, or application processes already in the system.

In an embodiment, nodes intercommunicate over a data fabric, such as the Schooner data fabric (SDF) or Sharing Data Fabric (discussed above). A data fabric is a unified user-space mechanism to access and store data into a hierarchical DRAM, flash memory, and storage sub-system of a clustered or distributed set of computer systems, and having user-defined attributes to control one or more of access, consistency, duplication, and durability.

Applications on processing nodes are enabled to access portions of the flash memory through node address spaces. The node address spaces can be thought of as reference spaces that application processes running (executing respective application program code) on various nodes in the cluster are enabled to use as a consistent way of referencing portions of the global distributed flash memory.

Additionally, the processing nodes are enabled to cache portions of the flash memory in other memory, such as DRAM in the memory address space of one or more of the application processes.

Application program code is enabled to access flash memory by performing data movement operations, for example SDP_GET and SDP_PUT. These operations in effect transfer portions of contents between flash memory and other memory, such as DRAM in the memory address space of application processes. Additionally, there are control operations which application programs use to determine when previous data operations have been performed in the system. SDP_SYNC is an example of such a control operation.

Techniques presented herein enable the implementation of a general set of data access primitives such as SDP_GET and SDP_PUT in a cluster of processing nodes and nodes containing flash memory. One skilled in the art will recognize that various other data access methods are also enabled by the techniques presented herein.

Application programs use a data Application Programming Interface (API) including, for example, SDP_GET and SDP_PUT, and various other data controls which are suitably enabled by techniques presented herein. One example embodiment of a data API is in a client library, which consists of software code that application programs link into their own code, and which includes application programming interface functions such as SOP_GET and SOP_PUT. Other embodiments are possible and do not limit the scope of the techniques presented herein.

Each node in the cluster runs one or more system processes that handle the processing of one or more of cross-node requests, inter-node communication, inter-node cache consistency, flash memory hardware access, flash memory wear management, data replication, and other cluster services and/or management functions. System processes optionally and/or selectively run in user mode as do application processes, and enable a logic layer, suitable to embody a high-level abstraction of the data in flash memory. The logic layer is usable by application programs. Techniques presented herein enable various abstractions, particularly suited for certain application programs, to be built in software on top of the basic data and communication mechanisms of the cluster. Other embodiments of system processes, such as co-location

25

of system process code with the application code in a single process model, are possible within the scope of the techniques presented herein.

The client library code is invoked on data access operations such as SOP_GET and SOP_PUT. The client library code is run entirely in user mode without any context switches or switches to kernel mode. In some usage scenarios, the API library immediately returns the data, such as when there is already a cached copy of the data at the processing node. If communication with other nodes is required as part of processing a data access operation, then a shared memory SCOREBOARD mechanism is used to invoke a system process. Various control and routing information is enabled to be passed through the scoreboard from the application process to the system process. Other embodiments of client library communication with a scoreboard, such as UNIX domain sockets, pipes, messages, or other well known mechanisms, may be employed within the scope of the techniques presented herein.

A system process is responsible for the cross-node processing of SDP_GET and SDP_PUT, and communicates with an invoking application process, or a software agent of the application process, through a shared memory scoreboard. An example processing flow for the SDP_GET operation is presented in FIG. 24, which is an illustration of a process flow of a SDP_GET according to an embodiment of the invention.

As illustrated in FIG. 23, SDP_GET is invoked by the application process (or a thread of the application process) when referencing a particular data element stored in flash memory. Thereafter, execution of SDP library code determines if the operation can be performed locally, for example when a cache hit is detected (step 1). If cross-node communication is needed to perform the operation, for example when a cache miss is detected, then the client library code creates a scoreboard entry and invokes the SDP process, which is an example of a system process.

As illustrated in FIG. 23, the system process examines the scoreboard entry and creates a system message to be sent to a system process on the node containing the flash memory or other data resource (step 3). The cluster network controller delivers the message to the node containing the data and a system process is invoked on that node to process the message (steps 5, 6). Next, the system process on the node containing the data performs the requested data access operation (steps 8-10) (for example by reading flash for SDP_GET), creates a response message containing the data and other control information, and sends it back to the node running the application process (step 11).

Again with reference to FIG. 23, the system process on the node running the application process receives the response message (step 12), deposits the data into a buffer in the application process memory address space (step 13), updates the scoreboard entry (step 14), and optionally signals the application process (step 15).

As shown in FIG. 23, the application process at this time considers the operation as having been performed and may proceed (step 16) (if it was synchronously waiting), or alternatively may, at any subsequent time, issue a SYNC on the particular operation which will succeed. The scoreboard entry, no longer needed, is freed for future use by a subsequent data access operation.

The process flow illustrated in FIG. 23 illustrates a cache miss. FIG. 22 is an illustration of a process flow for a cache hit at the action node according to an embodiment of the invention. As shown in FIG. 22, a worker thread of the application process at the action node initially consults the object cache. If a cache hit is made, then the data stored in the cache may be

26

obtained by the worker thread. Of course, if the desired data is not present in the object cache, then the process flow of FIG. 23 may be used to enable the worker thread to retrieve the desired data.

In the operation flow depicted in FIG. 23, the communication mechanism between application processes and the data fabric processes must be efficient as possible. Embodiments provide for a shared memory communication mechanism that allows a very efficient implementation and yet provides many additional features. Specifically, these features include metadata that tracks the progress of an operation as it is processed by the data fabric, DRAM buffers for data read from flash or written to flash, DRAM cache that is shared by one or more client processes on the same node, cluster membership information, mapping tables used to map logical object ID's to their physical location in terms of physical system, flash device, and location with a flash device, and buffering and metadata for the internode messaging system.

In an embodiment, the scoreboard may act as a communication mechanism. The communication mechanism of an embodiment is implemented on top of file-based shared memory that can be mapped in a process address space. The backing file can be placed on traditional persistent media, such as hard disks or flash memory, or may be on a RAM disk. One approach for creating this form of shared memory is using the "mmap" call available on most Unix operating systems. This communication mechanism provides a collection of methods for creating, accessing and manipulating shared memory by one or more application or system processes. The communication mechanism of an embodiment allows objects in a particular shared memory arena to be accessed by multiple user and/or system processes. Also, the communication mechanism of an embodiment provides for allocation and deallocation of shared memory objects by different processes and well as the dynamic addition/removal of client/system processes that have access to the shared memory. The communication mechanism of an embodiment also provides features beyond those provided by standard Unix shared memory facilities.

For example, the communication mechanism of an embodiment provides for reference counts are maintained per allocated object region so that multiple clients can reference the same memory region, and that region will not be deallocated until all clients have released it. As another example, the communication mechanism of an embodiment provides for specific regions of global shared memory to be marked as read-only to prevent data corruption and facilitate the detection of errors. Also, the communication mechanism of an embodiment provides a means of referencing an allocated region of shared memory (object) that can be passed between different user and system processes, and used as pointers within data structures kept within shared memory.

For maximum performance, communication mechanisms for an embodiment may allow shared memory regions to be marked as non-pageable.

A communication mechanism of an embodiment is crash tolerant, i.e., shared memory survives failures of client or system processes. This feature may be achieved by recording, for each allocated object, the identities of all processes that are currently accessing the object. This metadata is kept in the shared memory arenas, which survive process crashes. A shared memory manager process runs in the background and is responsible for handling process crashes. Advantageously, failure of one client on a node does not perturb activities of another client on the same node that doesn't fail. Also, if a system process crashes, then the process can be restarted and the state of shared memory prior to the crash may be recovered.

27

ered. This state of shared memory prior to the crash may provide information after the crash that can be used for debugging purposes. In an embodiment, if a client process crashes, then the shared memory manager updates all shared memory structures that were referenced by the now defunct client process.

The communication mechanism of an embodiment optionally allows an allocated region of shared memory to be asynchronously released by the shared memory manager, for example, if there is no free memory available and the memory manager needs memory for a high priority use, such as an incoming messaging buffer. This is useful for implementing and elastic object cache in which the amount of shared memory used for the cache can grow or shrink dynamically, based on how much memory is required for other SDF functions.

The communication mechanism of an embodiment supports multi-versioning, in which concurrent write operations may operate in parallel on distinct versions of the same object, improving performance. This may be achieved by providing an API (Application Programmer Interface) that allows an application to create multiple version and designate the current "permanent" version. Note that in this context the "application" is the Schooner Data Fabric (SDF) application library that implements the higher level get/set/etc. operations on top of this high function shared memory.

FIG. 20 is an illustration of a SDF Agent according to an embodiment of the invention. As depicted in FIG. 20, the SDF Agent on the action node is external to the application process. The SDF agent on the action node uses a shared scoreboard to communicate with a SDF process on the home node. GET/PUT requests are handled by the SDF process on the home node. SDF agent on the action node uses an interface to the application process thread schedule to signal application threads on sync completions.

FIG. 21 is an illustration of a process structure on the action node and home node according to an embodiment. As illustrated in FIG. 21, at the action node, API calls invoke the SDF library, which consults an object cache and post requests to the SDF scoreboard. The SDF agent thread sends and receives updates, updates the scoreboard, and activates worker threads on request completion.

As illustrated in FIG. 21, at the home node, the SDF processes use cooperative user level threads. These user level threads service incoming requests and switch on flash and third party cache requests. The scoreboard at the home node stores the state of outstanding requests. Background threads at the home node may perform performance related functionality, such as wear leveling.

FIG. 24 is an illustration of a scoreboard request block format according to an embodiment of the invention. The format depicted in FIG. 25 is merely illustrative of one embodiment, and is not meant to depict all embodiments of the invention.

FIGS. 25-31 illustrate an example of a scoreboard API and scoreboard interaction according to one embodiment of the invention. The scoreboard API and interaction illustrated in FIGS. 25-31 describe an approach for the allocation and tracking of shared memory that enables detection of and recovery from shared memory system errors such as a crashed application process that has outstanding buffers or scoreboard entries. Shared memory is further enabled to be managed as a system-wide resource that enables seamless system reconfiguration, allowing non-stop system maintenance operations and/or other advantages.

The communication mechanism of an embodiment provides synchronization and ordering constructs to facilitate

28

orderly concurrent access from multiple user/system processes. The state of the synchronization constructs is preserved in shared memory when a process crashes, so the global memory manager can clean up the synchronization state as needed (e.g., release locks that were held by a crashed process). Examples of such synchronization and ordering constructs are depicted in FIGS. 32-36 according to an embodiment.

One skilled in the art will recognize that various other data access primitives are also enabled by the techniques presented herein. Other data access primitives may include data transactions, bulk copies, and index lookups, and various other protocols.

Many embodiments are possible. Not all of these features need to be present in all embodiments, and many variations and sub-combinations of these features are contemplated by the inventor. The invention could be implemented in hardware, such as hardware logic gates, by a programmable processor either using firmware, software, or other code, or various combinations.

In some embodiments, the client library executes operations such as SDP_GET and SDP_PUT on behalf of an application process completely through the scoreboard, including operations that can be completed locally. This enables further decoupling of the application process from the system process.

In various embodiments, threads of the system process may be run inside the application process. This enables tighter coupling of the application process to the system process.

Values can be inverted, offset, combined with other values, and manipulated in many ways using known mathematical properties. An inversion could be added to an XOR to generate an exclusive-NOR (XNOR), but this is simply a derivative of an XOR and within a family of XOR functions. Other logic tricks and manipulations are contemplated and considered to be within the scope of the invention.

In the foregoing specification, embodiments of the invention have been described with reference to numerous specific details that may vary from implementation to implementation. Thus, the sole and exclusive indicator of what is the invention, and is intended by the applicants to be the invention, is the set of claims that issue from this application, in the specific form in which such claims issue, including any subsequent correction. Any definitions expressly set forth herein for terms contained in such claims shall govern the meaning of such terms as used in the claims. Hence, no limitation, element, property, feature, advantage or attribute that is not expressly recited in a claim should limit the scope of such claim in any way. The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.

What is claimed is:

1. A distributed storage system, comprising:

a plurality of nodes, wherein each node of the plurality of nodes executes one or more application processes that access persistent shared memory, wherein each of the one or more application processes is maintained on a respective node of the plurality of nodes;

the persistent shared memory, wherein the persistent shared memory is implemented by solid state devices maintained on the plurality of nodes;

a scoreboard implemented in the persistent shared memory, the scoreboard storing one or more scoreboard entries corresponding to a respective outstanding data

29

access operation for accessing data stored in the persistent shared memory, each of the one or more scoreboard entries including:

- data identifying a respective outstanding data access operation;
- a thread ID identifying a thread, of the one or more application processes, that initiated the respective outstanding data access operation; and
- a completion status associated with the respective outstanding data access operation,

a shared data fabric to enable the one or more application processes to access the persistent shared memory; and

a process in a first node of the plurality of nodes to update a first scoreboard entry of the one or more scoreboard entries, the first scoreboard entry corresponding to a first data access operation performed by an application process executed by the first node.

2. The distributed storage system of claim 1, wherein regions of the persistent shared memory are allocated by one process and deallocated by a different process.

3. The distributed storage system of claim 1, wherein reference counts are maintained for regions of the persistent shared memory, and wherein only unreferenced regions of the persistent shared memory may be deallocated.

4. The distributed storage system of claim 1, further comprising a communication mechanism to designate one or more regions of the persistent shared memory as non-pageable.

5. The distributed storage system of claim 1, including:

- a process in the first node of the plurality of nodes to free the first scoreboard entry for use by a subsequent data access operation, wherein freeing the first scoreboard entry is in accordance with completion of the first data access operation.

6. A method for accessing data in a distributed storage system, the method comprising:

- at a first node in a distributed storage system comprising a plurality of nodes, performing operations comprising:
 - executing one or more application processes that access persistent shared memory, wherein:
 - executing the one or more application processes includes performing a first data access operation,
 - the persistent shared memory is implemented by solid state devices maintained on the plurality of nodes, and
- a shared data fabric enables the one or more application processes to access the persistent shared memory;

30

determining whether a data object is in a memory of the node;

in accordance with a determination that the data object is not in the memory of the node:

- creating a first scoreboard entry to be stored in a scoreboard implemented in the persistent shared memory, wherein the scoreboard stores one or more scoreboard entries, including the first scoreboard entry, corresponding to a respective outstanding data access operation for accessing data stored in the persistent shared memory, and wherein the first scoreboard entry corresponds to the first data access operation, each of the one or more scoreboard entries including:
 - data identifying a respective outstanding data access operation;
 - a thread ID identifying a thread, of the one or more application processes, that initiated the respective outstanding data access operation; and
 - a completion status associated with the respective outstanding data access operation; and
- updating the first scoreboard entry.

7. The method of claim 6, wherein regions of the persistent shared memory are allocated by one process and deallocated by a different process.

8. The method of claim 6, wherein reference counts are maintained for regions of the persistent shared memory, and wherein only unreferenced regions of the persistent shared memory may be deallocated.

9. The method of claim 6, wherein one or more regions of the persistent shared memory is designated as non-pageable.

10. The method of claim 6, further comprising:

- creating a system message based on the first scoreboard entry;
- sending, to a second node of the plurality of nodes, the system message; and
- receiving, from the second node of the plurality of nodes, a response message containing the data object, wherein updating the first scoreboard entry includes updating the respective completion status associated with the first data access operation.

11. The method of claim 6, further comprising:

- in accordance with completion of the first data access operation, freeing the first scoreboard entry for use by a subsequent data access operation.

* * * * *